Knowledge-Aware Reasoning over Multimodal Semi-structured Tables

Suyash Vardhan Mathur[1], Jainit Bafna[1]*, Kunal Kartik[2]*, Harshita Khandelwal[3]*
Manish Shrivastava[1], Vivek Gupta[4], Mohit Bansal[5], Dan Roth[4]

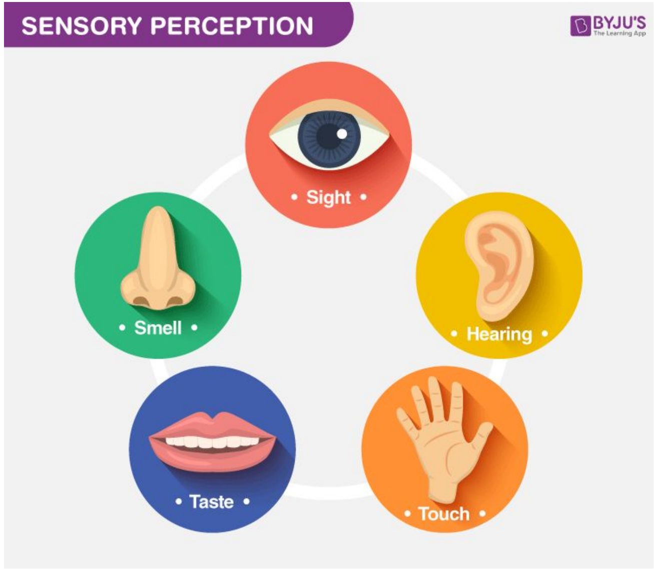
[1]IIIT Hyderabad, [2]IIT Guwahati, [3]UCLA, [4]UPenn, [5]UNC Chapel Hill

(Accepted to EMNLP'24)

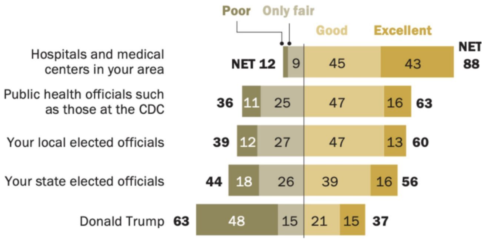
Truly Intelligent Machines are Multimodal

Humans can process information from **sight, hearing, smell, taste,** and **touch**.

Processing information from **multiple** modalities is crucial for AI to mimic human intelligence.



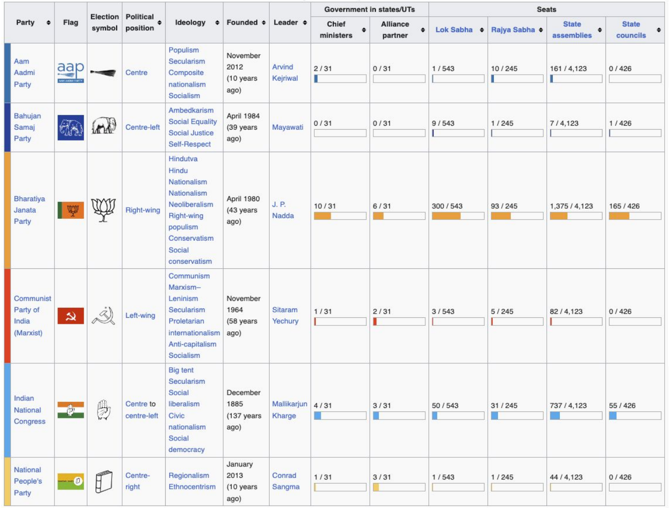
STRUCTURAL MULTIMODAL DATA



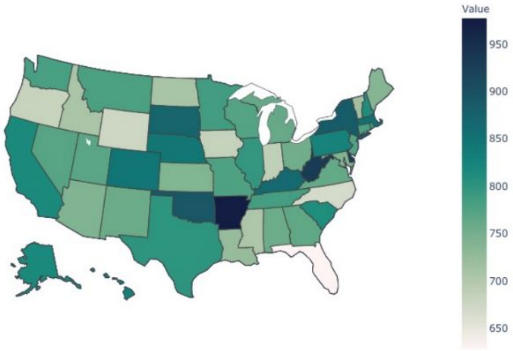
Majority of Americans are critical of Trump's response to COVID-19; nearly half say he is doing 'poor' job

% who rate the job each of the following is doing responding to the coronavirus outbreak as …

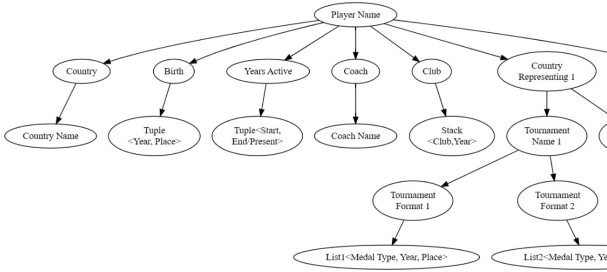
Note: No answer responses not shown.
Source: Survey of U.S. adults conducted July 27-Aug. 2, 2020.

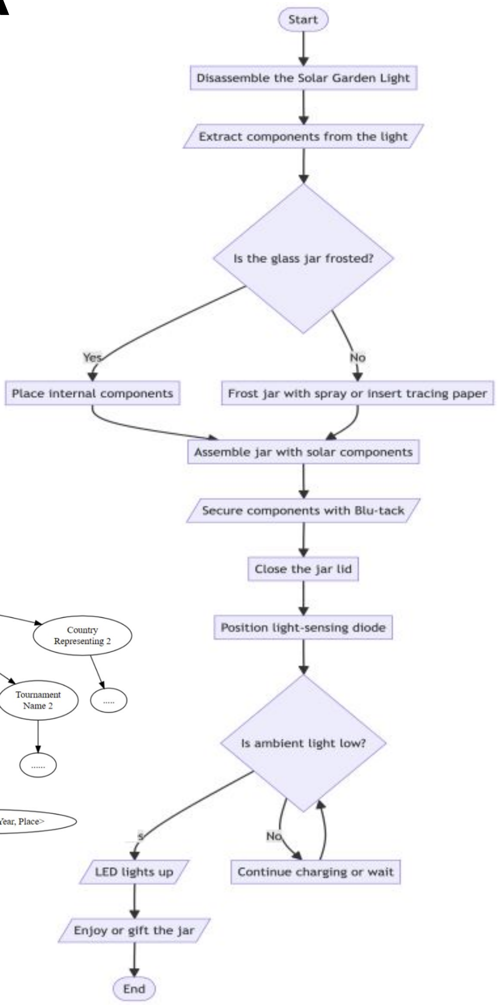


Party	Flag	Election symbol	Political position	Ideology	Founded	Leader	Government in states/UTs		Seats			
							Chief ministers	Alliance partner	Lok Sabha	Rajya Sabha	State assemblies	State councils
Aam Aadmi Party			Centre	Populism Secularism Composite nationalism Socialism	November 2012 (10 years ago)	Arvind Kejriwal	2 / 31	0 / 31	1 / 543	10 / 245	161 / 4,123	0 / 426
Bahujan Samaj Party			Centre-left	Ambedkarism Social Equality Social Justice Self-Respect	April 1984 (39 years ago)	Mayawati	0 / 31	0 / 31	9 / 543	1 / 245	7 / 4,123	1 / 426
Bharatiya Janata Party			Right-wing	Hindutva Hindu Nationalism Nationalism Neoliberalism Right-wing populism Conservatism Social conservatism	April 1980 (43 years ago)	J. P. Nadda	10 / 31	6 / 31	300 / 543	93 / 245	1,375 / 4,123	165 / 426
Communist Party of India (Marxist)			Left-wing	Communism Marxism–Leninism Secularism Proletarian internationalism Anti-capitalism Socialism	November 1964 (58 years ago)	Sitaram Yechury	1 / 31	2 / 31	3 / 543	5 / 245	82 / 4,123	0 / 426
Indian National Congress			Centre to centre-left	Big tent Secularism Social liberalism Civic nationalism Social democracy	December 1885 (137 years ago)	Mallikarjun Kharge	4 / 31	3 / 31	50 / 543	31 / 245	737 / 4,123	55 / 426
National People's Party			Centre-right	Regionalism Ethnocentrism	January 2013 (10 years ago)	Conrad Sangma	1 / 31	3 / 31	1 / 543	1 / 245	44 / 4,123	0 / 426



Q: Among the regions that border Texas, which have the lowest value?
A: [Louisiana]



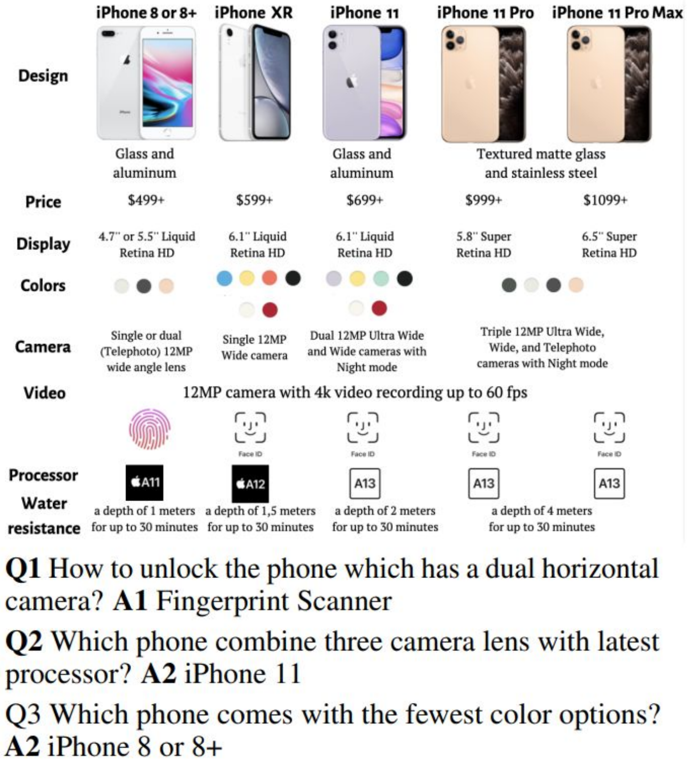

Start
Disassemble the Solar Garden Light
Extract components from the light
Is the glass jar frosted?
Yes
No
Place internal components
Frost jar with spray or insert tracing paper
Assemble jar with solar components
Secure components with Blu-tack
Close the jar lid
Position light-sensing diode
Is ambient light low?
s
No
LED lights up
Continue charging or wait
Enjoy or gift the jar
End

Problem Statement

Tables **summarize information** efficiently, but NLP research mostly focused on text-only tables.

Real-world tables often **include images** which are representing entities, such as:

- Products in e-commerce
- Team logos in sports
- Visual symptoms in medical comparisons
- Molecular diagrams in educational content
- Political Party Symbols
- and more…



	iPhone 8 or 8+	iPhone XR	iPhone 11	iPhone 11 Pro	iPhone 11 Pro Max
Design	Glass and aluminum		Glass and aluminum	Textured matte glass and stainless steel	
Price	$499+	$599+	$699+	$999+	$1099+
Display	4.7" or 5.5" Liquid Retina HD	6.1" Liquid Retina HD	6.1" Liquid Retina HD	5.8" Super Retina HD	6.5" Super Retina HD
Colors	● ● ●	● ● ● ● ● ●	● ● ● ● ● ●	● ● ● ●	
Camera	Single or dual (Telephoto) 12MP wide angle lens	Single 12MP Wide camera	Dual 12MP Ultra Wide and Wide cameras with Night mode	Triple 12MP Ultra Wide, Wide, and Telephoto cameras with Night mode	
Video	12MP camera with 4k video recording up to 60 fps				
		Face ID	Face ID	Face ID	Face ID
Processor	A11	A12	A13	A13	A13
Water resistance	a depth of 1 meters for up to 30 minutes	a depth of 1,5 meters for up to 30 minutes	a depth of 2 meters for up to 30 minutes	a depth of 4 meters for up to 30 minutes	

Q1 How to unlock the phone which has a dual horizontal camera? **A1** Fingerprint Scanner

Q2 Which phone combine three camera lens with latest processor? **A2** iPhone 11

Q3 Which phone comes with the fewest color options? **A2** iPhone 8 or 8+

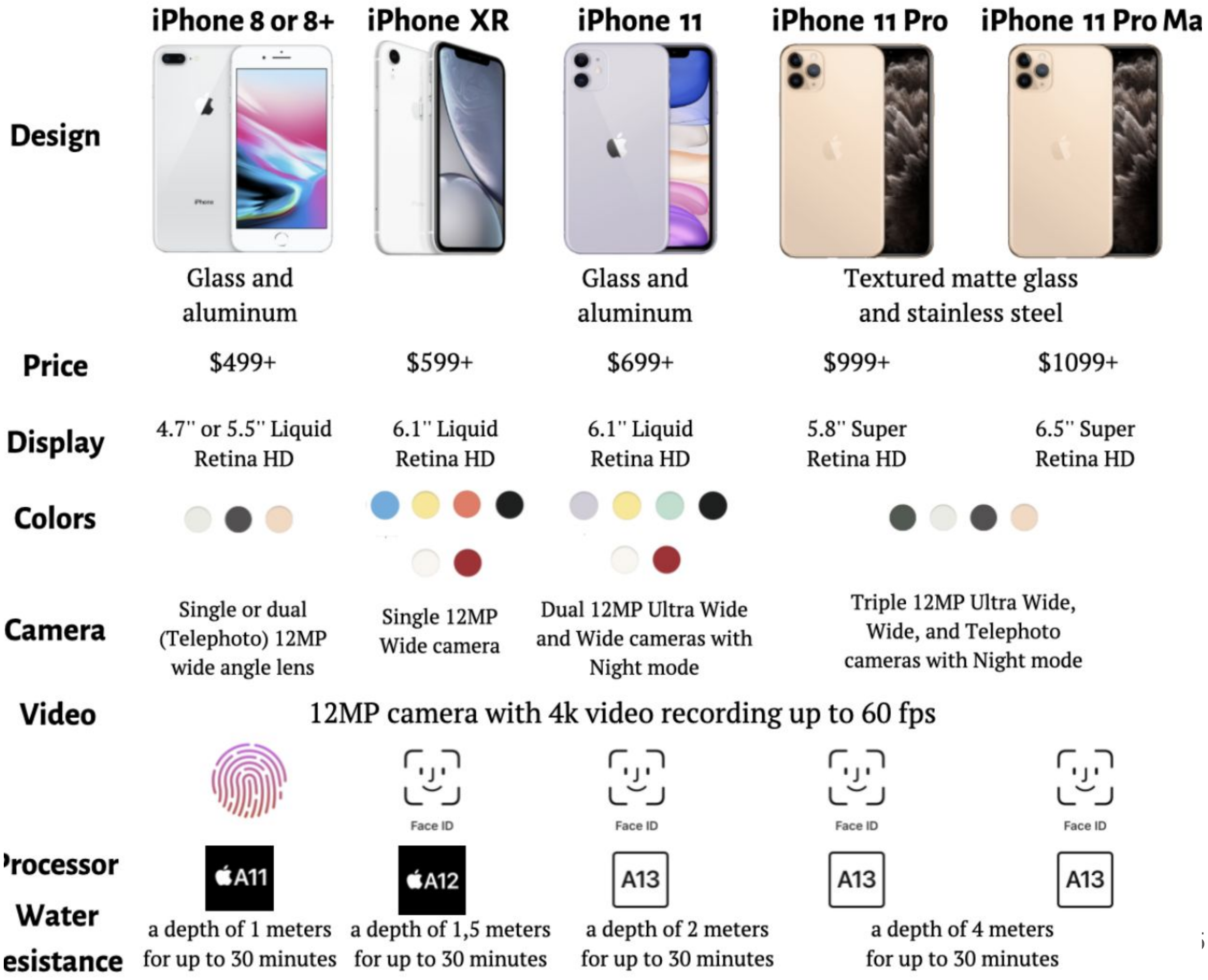
MultiModal Tables

Visual Entity Linking

- especially **implicit QA** image → text

Visual: Symbols ($, face, A12), Colors, Patterns, infographic, Charts etc.

Structure: Hierarchical Tables/Lists, Hyperlink, Sub-header



	iPhone 8 or 8+	iPhone XR	iPhone 11	iPhone 11 Pro	iPhone 11 Pro Ma
Design	Glass and aluminum		Glass and aluminum	Textured matte glass and stainless steel	
Price	$499+	$599+	$699+	$999+	$1099+
Display	4.7" or 5.5" Liquid Retina HD	6.1" Liquid Retina HD	6.1" Liquid Retina HD	5.8" Super Retina HD	6.5" Super Retina HD
Colors	● ● ●	● ● ● ● ● ●	● ● ● ● ● ●	● ● ● ●	
Camera	Single or dual (Telephoto) 12MP wide angle lens	Single 12MP Wide camera	Dual 12MP Ultra Wide and Wide cameras with Night mode	Triple 12MP Ultra Wide, Wide, and Telephoto cameras with Night mode	
Video	12MP camera with 4k video recording up to 60 fps				
		Face ID	Face ID	Face ID	Face ID
rocessor	A11	A12	A13	A13	A13
Water esistance	a depth of 1 meters for up to 30 minutes	a depth of 1,5 meters for up to 30 minutes	a depth of 2 meters for up to 30 minutes	a depth of 4 meters for up to 30 minutes	

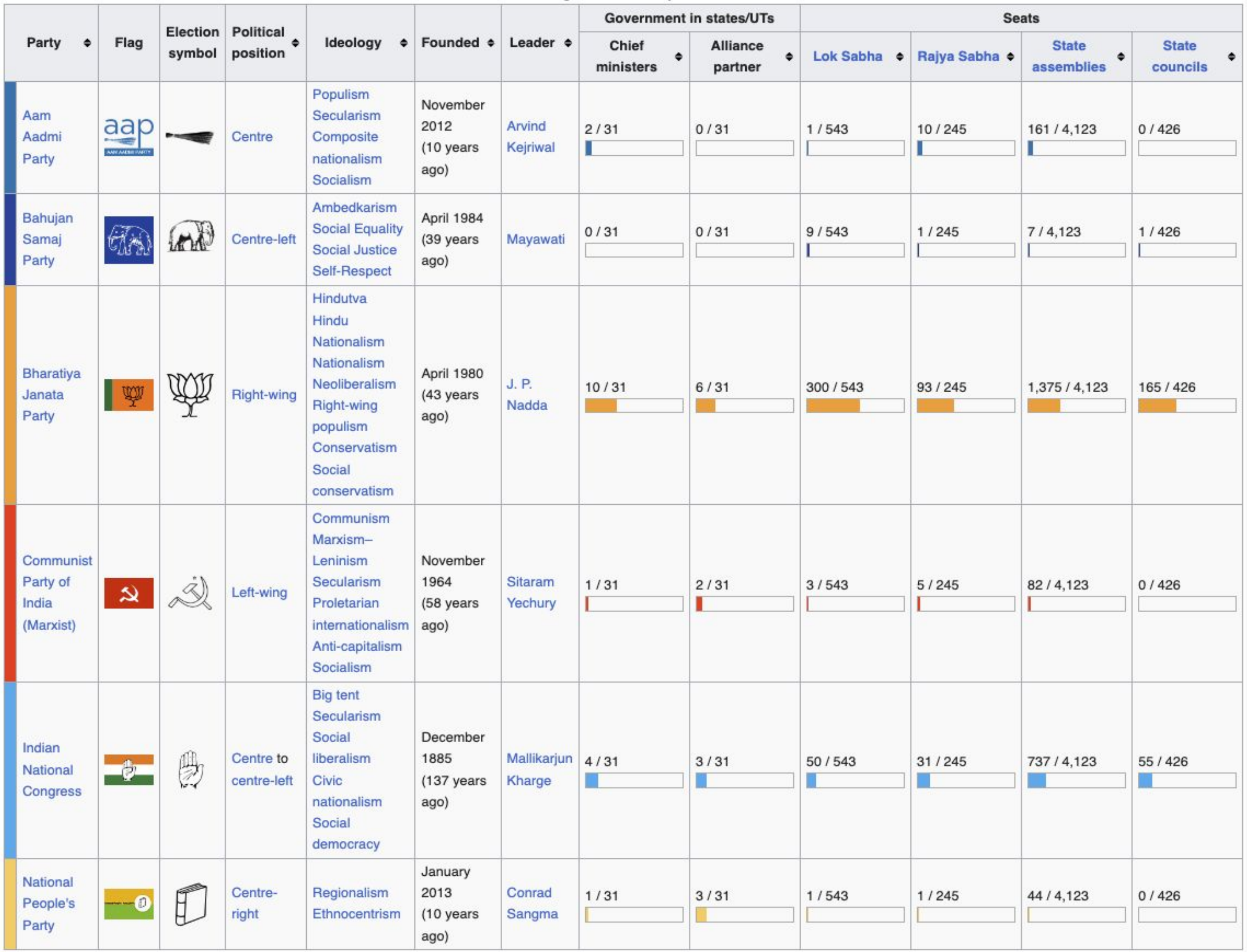
Visual Questions

(even more challenging)

Q. How many states did party with lotus symbol govern?
A. 16

Q. How many of these are with an Alliance partners.
A. 6

Q. Which is the oldest party election symbol?
A. Hand



Party	Flag	Election symbol	Political position	Ideology	Founded	Leader	Government in states/UTs		Seats			
							Chief ministers	Alliance partner	Lok Sabha	Rajya Sabha	State assemblies	State councils
Aam Aadmi Party			Centre	Populism Secularism Composite nationalism Socialism	November 2012 (10 years ago)	Arvind Kejriwal	2 / 31	0 / 31	1 / 543	10 / 245	161 / 4,123	0 / 426
Bahujan Samaj Party			Centre-left	Ambedkarism Social Equality Social Justice Self-Respect	April 1984 (39 years ago)	Mayawati	0 / 31	0 / 31	9 / 543	1 / 245	7 / 4,123	1 / 426
Bharatiya Janata Party			Right-wing	Hindutva Hindu Nationalism Nationalism Neoliberalism Right-wing populism Conservatism Social conservatism	April 1980 (43 years ago)	J. P. Nadda	10 / 31	6 / 31	300 / 543	93 / 245	1,375 / 4,123	165 / 426
Communist Party of India (Marxist)			Left-wing	Communism Marxism–Leninism Secularism Proletarian internationalism Anti-capitalism Socialism	November 1964 (58 years ago)	Sitaram Yechury	1 / 31	2 / 31	3 / 543	5 / 245	82 / 4,123	0 / 426
Indian National Congress			Centre to centre-left	Big tent Secularism Social liberalism Civic nationalism Social democracy	December 1885 (137 years ago)	Mallikarjun Kharge	4 / 31	3 / 31	50 / 543	31 / 245	737 / 4,123	55 / 426
National People's Party			Centre-right	Regionalism Ethnocentrism	January 2013 (10 years ago)	Conrad Sangma	1 / 31	3 / 31	1 / 543	1 / 245	44 / 4,123	0 / 426

Complex Question (Images + Charts)

Which team is ranked 5th and what is their points percentage?

Identify the team with the most number of draws.

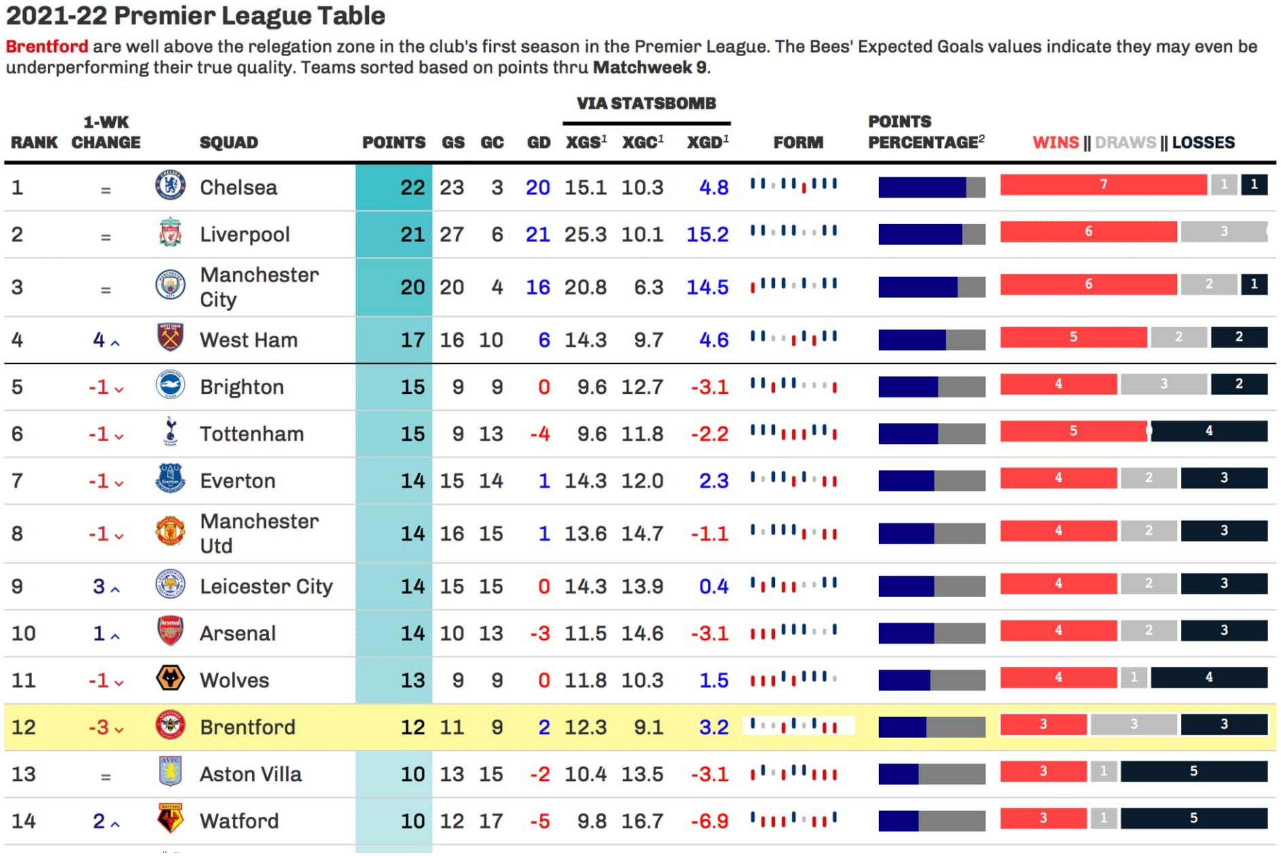
How many wins does Tottenham have?

Identify the team with 4 wins and a points percentage close to 30%.

Team with blue lion in its logo, has the highest goal difference of how much?

Team with a cannon logo has 14 points and a goal difference of -3. What is their current rank?

team with a lion has 14 points, scored 14 goals, conceded with how many points?

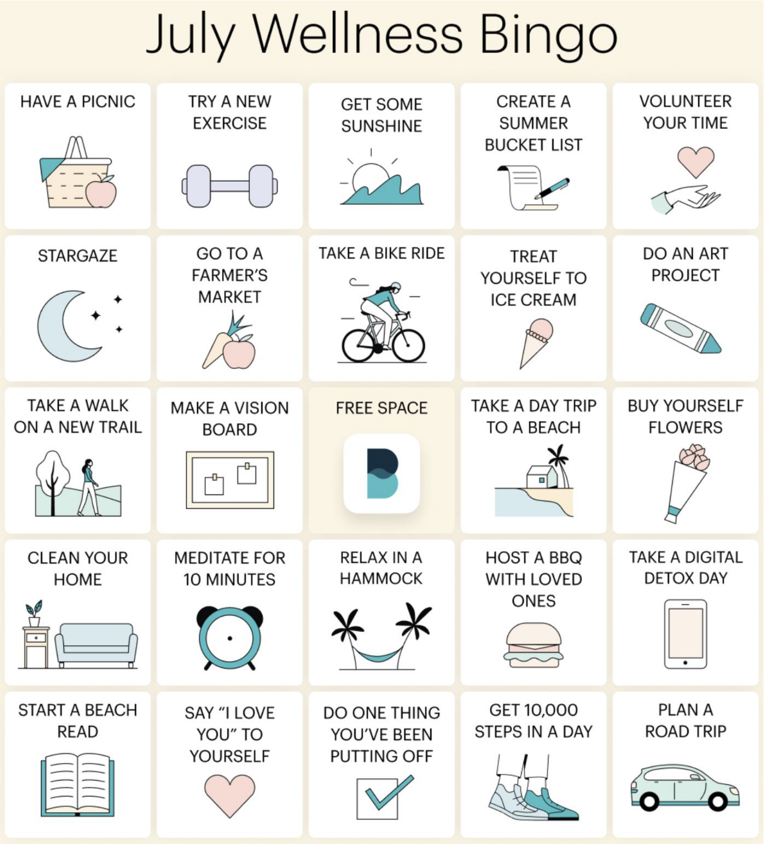


2021-22 Premier League Table

Brentford are well above the relegation zone in the club's first season in the Premier League. The Bees' Expected Goals values indicate they may even be underperforming their true quality. Teams sorted based on points thru **Matchweek 9**.

							VIA STATSBOMB					
RANK	1-WK CHANGE	SQUAD	POINTS	GS	GC	GD	XGS[1]	XGC[1]	XGD[1]	FORM	POINTS PERCENTAGE[2]	WINS ‖ DRAWS ‖ LOSSES
1	=	Chelsea	22	23	3	20	15.1	10.3	4.8			7 ‖ 1 ‖ 1
2	=	Liverpool	21	27	6	21	25.3	10.1	15.2			6 ‖ 3
3	=	Manchester City	20	20	4	16	20.8	6.3	14.5			6 ‖ 2 ‖ 1
4	4 ^	West Ham	17	16	10	6	14.3	9.7	4.6			5 ‖ 2 ‖ 2
5	-1 ⌄	Brighton	15	9	9	0	9.6	12.7	-3.1			4 ‖ 3 ‖ 2
6	-1 ⌄	Tottenham	15	9	13	-4	9.6	11.8	-2.2			5 ‖ 4
7	-1 ⌄	Everton	14	15	14	1	14.3	12.0	2.3			4 ‖ 2 ‖ 3
8	-1 ⌄	Manchester Utd	14	16	15	1	13.6	14.7	-1.1			4 ‖ 2 ‖ 3
9	3 ^	Leicester City	14	15	15	0	14.3	13.9	0.4			4 ‖ 2 ‖ 3
10	1 ^	Arsenal	14	10	13	-3	11.5	14.6	-3.1			4 ‖ 2 ‖ 3
11	-1 ⌄	Wolves	13	9	9	0	11.8	10.3	1.5			4 ‖ 1 ‖ 4
12	-3 ⌄	Brentford	12	11	9	2	12.3	9.1	3.2			3 ‖ 3 ‖ 3
13	=	Aston Villa	10	13	15	-2	10.4	13.5	-3.1			3 ‖ 1 ‖ 5
14	2 ^	Watford	10	12	17	-5	9.8	16.7	-6.9			3 ‖ 1 ‖ 5

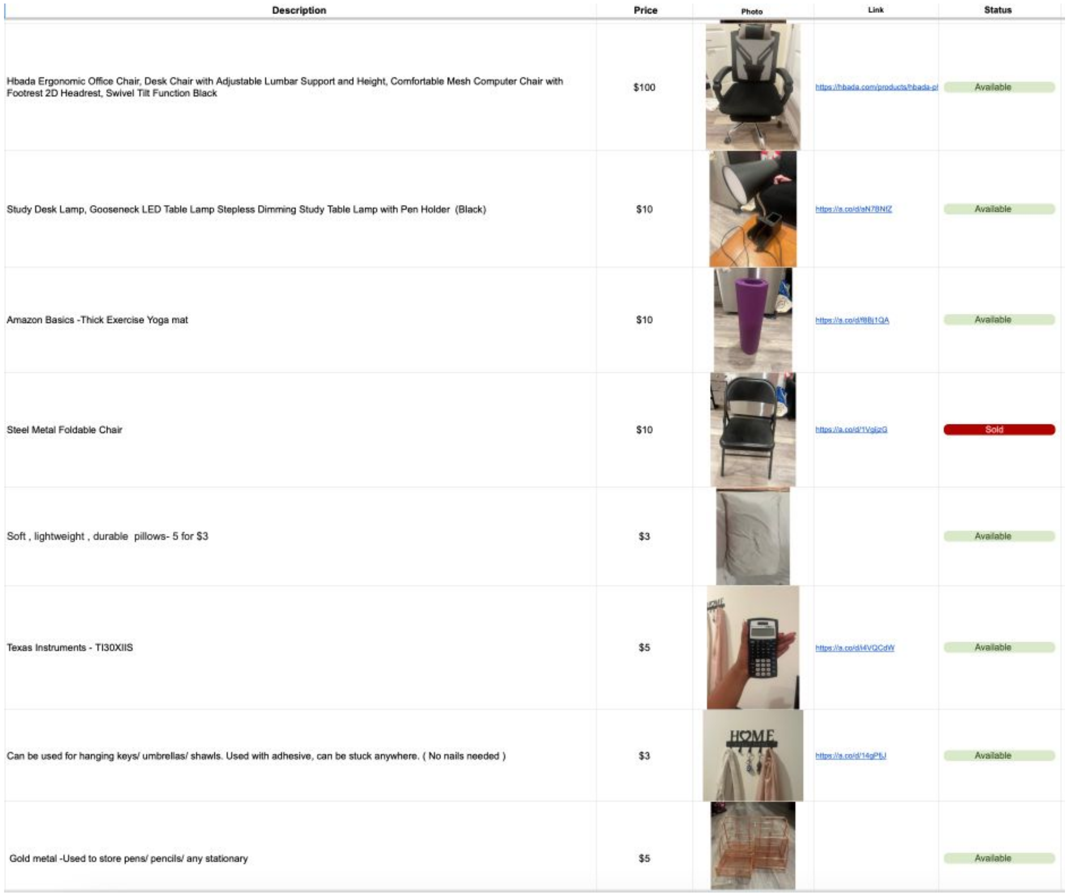
More Examples



July Wellness Bingo

HAVE A PICNIC	TRY A NEW EXERCISE	GET SOME SUNSHINE	CREATE A SUMMER BUCKET LIST	VOLUNTEER YOUR TIME
STARGAZE	GO TO A FARMER'S MARKET	TAKE A BIKE RIDE	TREAT YOURSELF TO ICE CREAM	DO AN ART PROJECT
TAKE A WALK ON A NEW TRAIL	MAKE A VISION BOARD	FREE SPACE	TAKE A DAY TRIP TO A BEACH	BUY YOURSELF FLOWERS
CLEAN YOUR HOME	MEDITATE FOR 10 MINUTES	RELAX IN A HAMMOCK	HOST A BBQ WITH LOVED ONES	TAKE A DIGITAL DETOX DAY
START A BEACH READ	SAY "I LOVE YOU" TO YOURSELF	DO ONE THING YOU'VE BEEN PUTTING OFF	GET 10,000 STEPS IN A DAY	PLAN A ROAD TRIP

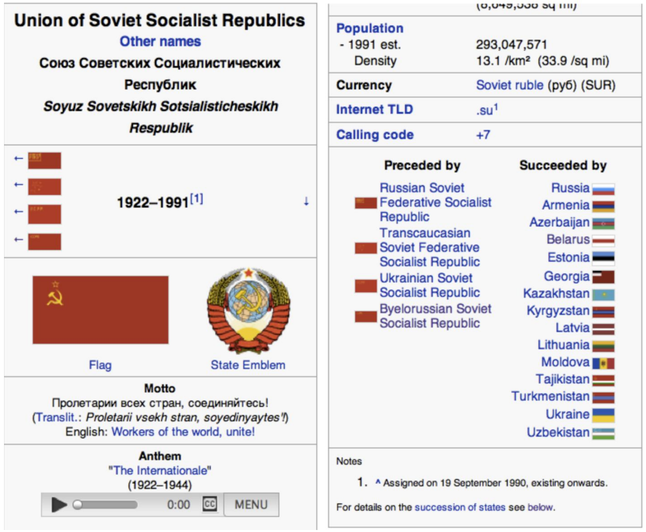
Sports: Symbols



Description	Price	Photo	Link	Status
Hbada Ergonomic Office Chair, Desk Chair with Adjustable Lumbar Support and Height, Comfortable Mesh Computer Chair with Footrest 2D Headrest, Swivel Tilt Function Black	$100		https://hbada.com/products/hbada-p/	Available
Study Desk Lamp, Gooseneck LED Table Lamp Stepless Dimming Study Table Lamp with Pen Holder (Black)	$10		https://a.co/d/aN7BNfZ	Available
Amazon Basics -Thick Exercise Yoga mat	$10		https://a.co/d/f8Bj1QA	Available
Steel Metal Foldable Chair	$10		https://a.co/d/1VglzG	Sold
Soft , lightweight , durable pillows- 5 for $3	$3			Available
Texas Instruments - TI30XIIS	$5		https://a.co/d/4VQCdW	Available
Can be used for hanging keys/ umbrellas/ shawls. Used with adhesive, can be stuck anywhere. (No nails needed)	$3		https://a.co/d/14gPfJ	Available
Gold metal -Used to store pens/ pencils/ any stationary	$5			Available

E-commerce: selling list

More Examples



Union of Soviet Socialist Republics
Other names
Союз Советских Социалистических Республик
Soyuz Sovetskikh Sotsialisticheskikh Respublik

1922–1991[1]

Motto
Пролетарии всех стран, соединяйтесь!
(Translit.: *Proletarii vsekh stran, soyedinyaytes'!*)
English: Workers of the world, unite!

Anthem
"The Internationale"
(1922–1944)

Population	
- 1991 est.	293,047,571
Density	13.1 /km² (33.9 /sq mi)
Currency	Soviet ruble (руб) (SUR)
Internet TLD	.su[1]
Calling code	+7

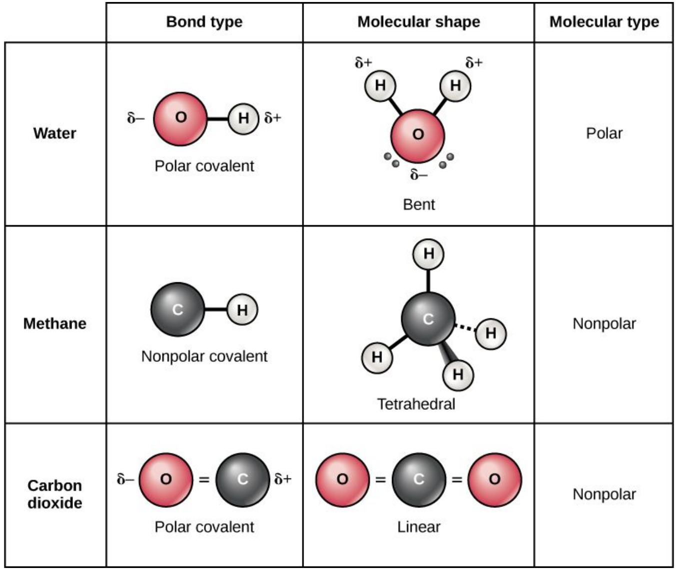
Preceded by	Succeeded by
Russian Soviet Federative Socialist Republic	Russia
Transcaucasian Soviet Federative Socialist Republic	Armenia
Ukrainian Soviet Socialist Republic	Azerbaijan
Byelorussian Soviet Socialist Republic	Belarus
	Estonia
	Georgia
	Kazakhstan
	Kyrgyzstan
	Latvia
	Lithuania
	Moldova
	Tajikistan
	Turkmenistan
	Ukraine
	Uzbekistan

Notes

1. ^ Assigned on 19 September 1990, existing onwards.

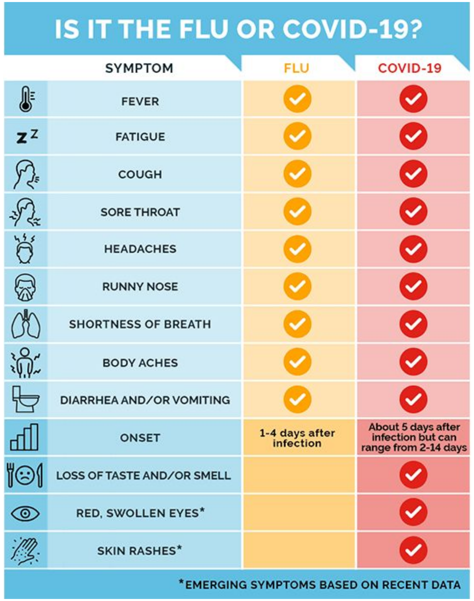
For details on the succession of states see below.

Conflict and Politics: Logos



	Bond type	Molecular shape	Molecular type
Water	Polar covalent	Bent	Polar
Methane	Nonpolar covalent	Tetrahedral	Nonpolar
Carbon dioxide	Polar covalent	Linear	Nonpolar

Chemistry: Molecular Diagrams

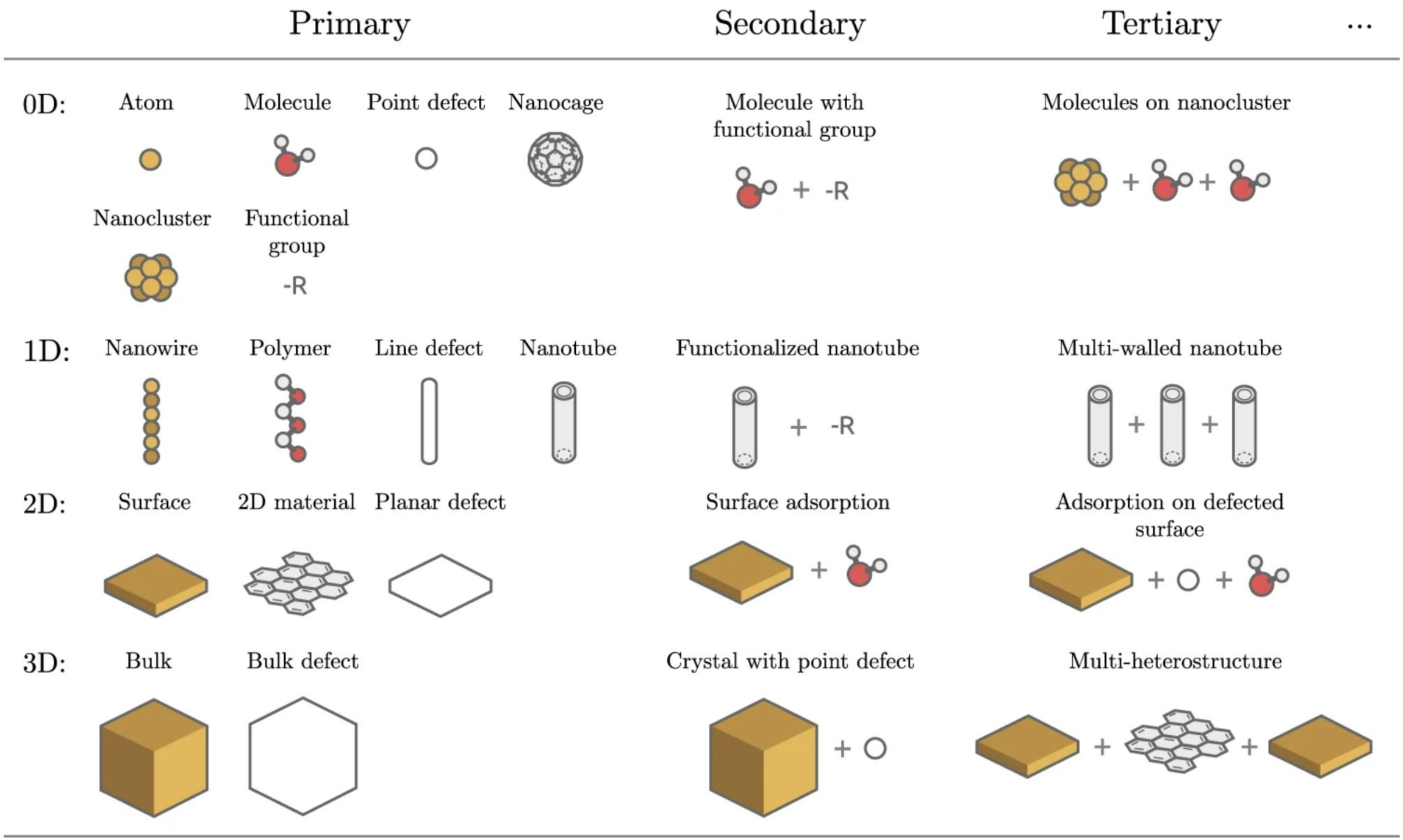


IS IT THE FLU OR COVID-19?

SYMPTOM	FLU	COVID-19
FEVER	✓	✓
FATIGUE	✓	✓
COUGH	✓	✓
SORE THROAT	✓	✓
HEADACHES	✓	✓
RUNNY NOSE	✓	✓
SHORTNESS OF BREATH	✓	✓
BODY ACHES	✓	✓
DIARRHEA AND/OR VOMITING	✓	✓
ONSET	1-4 days after infection	About 5 days after infection but can range from 2-14 days
LOSS OF TASTE AND/OR SMELL		✓
RED, SWOLLEN EYES*		✓
SKIN RASHES*		✓

*EMERGING SYMPTOMS BASED ON RECENT DATA

Medical: Symptoms


Primary
Secondary
Tertiary
...
0D:
Atom
Molecule
Point defect
Nanocage
Nanocluster
Functional group
-R
Molecule with functional group
+ -R
Molecules on nanocluster
+
+
1D:
Nanowire
Polymer
Line defect
Nanotube
Functionalized nanotube
+ -R
Multi-walled nanotube
+
+
2D:
Surface
2D material
Planar defect
Surface adsorption
+
Adsorption on defected surface
+
+
3D:
Bulk
Bulk defect
Crystal with point defect
+
Multi-heterostructure
+
+

Challenges in Multimodal Tabular Reasoning

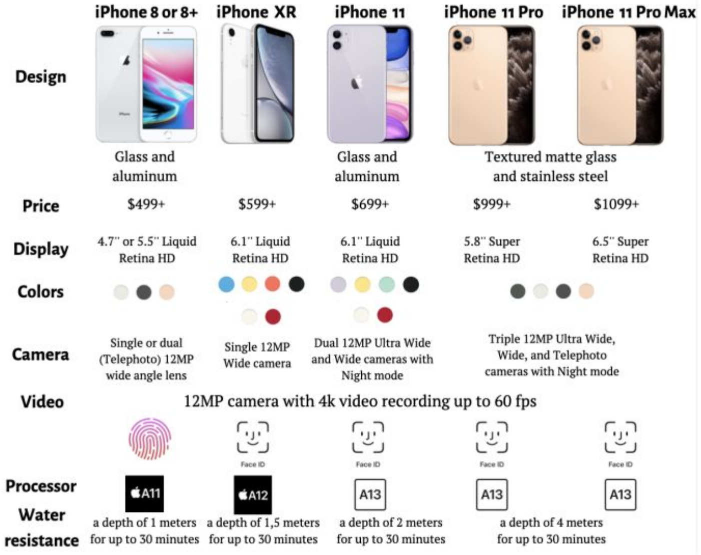
Entity Disambiguation: Identifying table entities for images using rows, headers, and other images as context.

Logical Reasoning: Applying complex reasoning, including numerical, temporal, commonsense, and entity-specific domain knowledge.

Multiple Images: Most VLMs handle single images; managing multiple structured images is challenging.

Contextual Understanding: Integrating information from various table parts and images for accurate interpretation.

Scalability Issues: Analyzing large tables with many images is resource-intensive and complex.



	iPhone 8 or 8+	iPhone XR	iPhone 11	iPhone 11 Pro	iPhone 11 Pro Max
Design	Glass and aluminum		Glass and aluminum	Textured matte glass and stainless steel	
Price	$499+	$599+	$699+	$999+	$1099+
Display	4.7" or 5.5" Liquid Retina HD	6.1" Liquid Retina HD	6.1" Liquid Retina HD	5.8" Super Retina HD	6.5" Super Retina HD
Colors	● ● ●	● ● ● ● ● ●	● ● ● ● ● ●	● ● ● ●	
Camera	Single or dual (Telephoto) 12MP wide angle lens	Single 12MP Wide camera	Dual 12MP Ultra Wide and Wide cameras with Night mode	Triple 12MP Ultra Wide, Wide, and Telephoto cameras with Night mode	
Video	12MP camera with 4k video recording up to 60 fps				
		Face ID	Face ID	Face ID	Face ID
Processor	A11	A12	A13	A13	A13
Water resistance	a depth of 1 meters for up to 30 minutes	a depth of 1,5 meters for up to 30 minutes	a depth of 2 meters for up to 30 minutes	a depth of 4 meters for up to 30 minutes	

Q1 How to unlock the phone which has a dual horizontal camera? **A1** Fingerprint Scanner

Q2 Which phone combine three camera lens with latest processor? **A2** iPhone 11

Q3 Which phone comes with the fewest color options? **A2** iPhone 8 or 8+

Dataset Recasting Pipeline

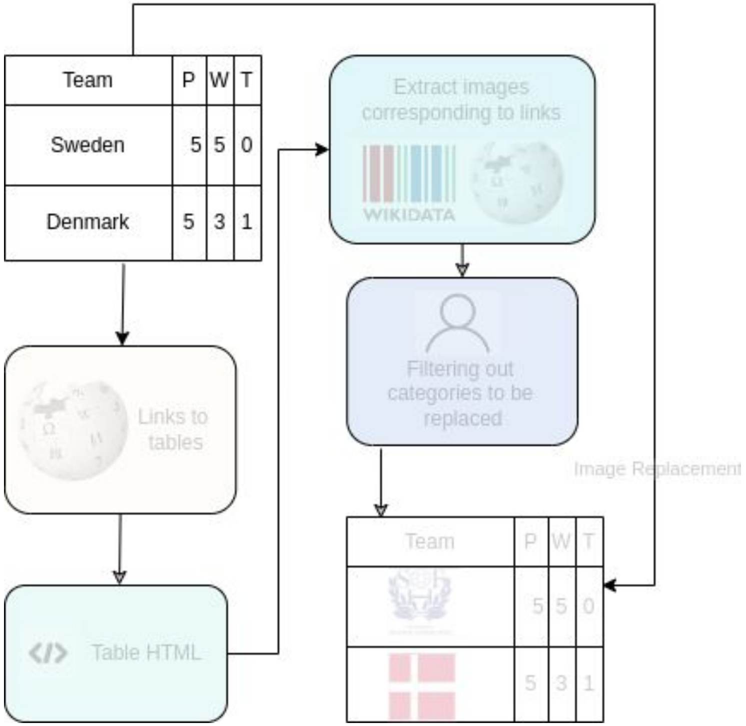
We **repurpose** four existing Wikipedia Table QA Datasets:

WikiSQL benchmark model capabilities in parsing entities accurately to answer SQL-based questions.

WikiTableQuestions include questions which require more complex reasoning.

FeTaQA include long-form answer questions which involve multiple row/column reasoning.

HybridQA includes contextual passages beyond tables, which require hybrid complex reasoning



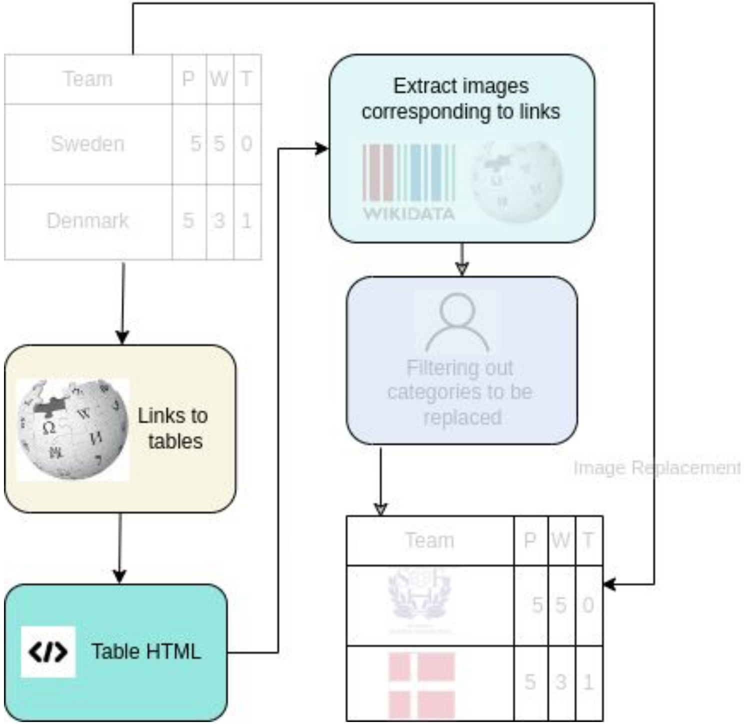
Dataset Recasting Pipeline

Converting a text table to multimodal requires mapping entity mentions to images.

Since a Wikipedia URL points to a unique infobox/Wikidata image, we can map

entity mentions → Wiki URL → Wiki images

We use page links to scrape raw HTML and apply the Jaccard Coefficient to locate tables.



Team	P	W	T
Sweden	5	5	0
Denmark	5	3	1

Links to tables

Table HTML

Extract images corresponding to links

Filtering out categories to be replaced

Image Replacement

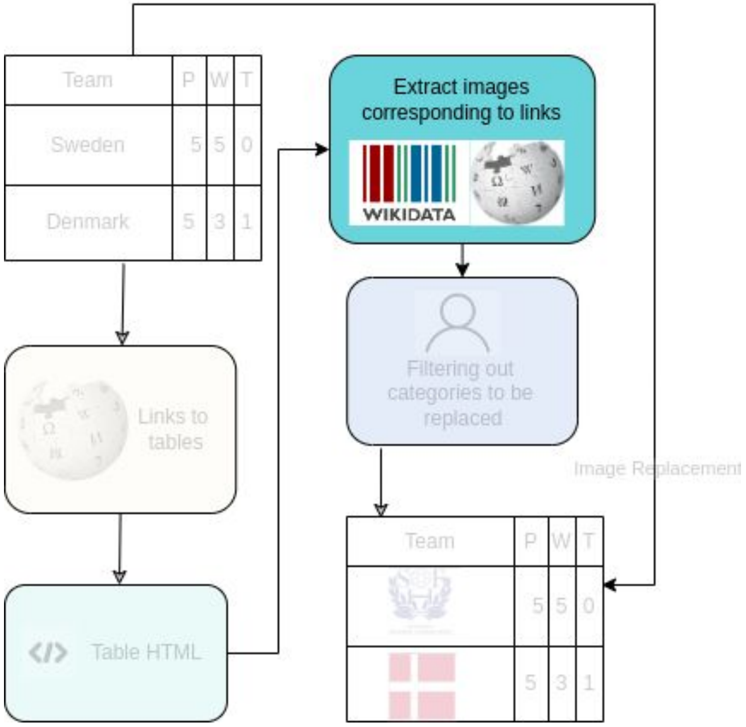
Team	P	W	T
	5	5	0
	5	3	1

Dataset Recasting Pipeline

We extract the Wikipedia links from the HTML and find their corresponding images.

We prioritize Wikipedia Infobox images and using a defined

- priority order for Wikidata images
- *Our order: image*, *logo image*, *flag image*, *coat of arms image*, *service ribbon image*.



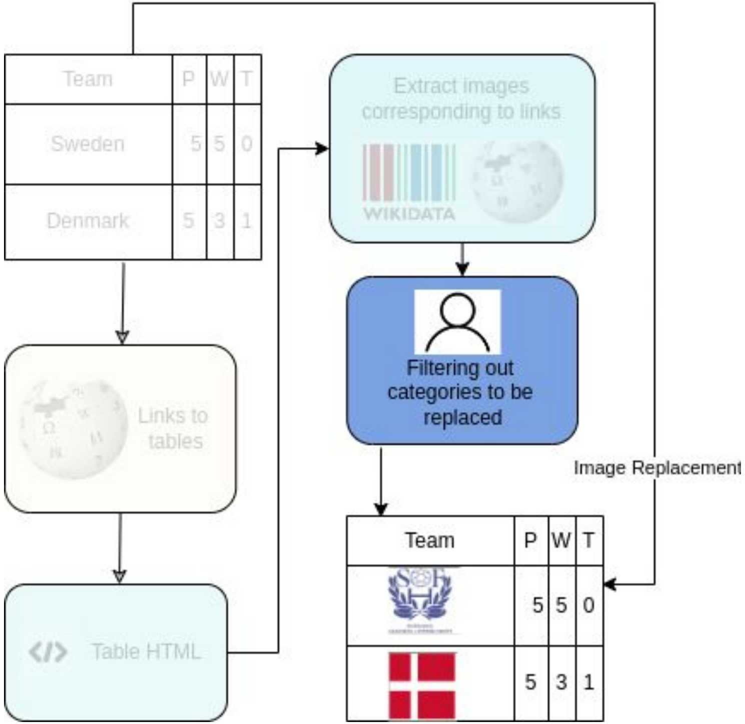
Dataset Recasting Pipeline

We **filter entities** for img replacement based on

- wikidata P31 *"instance of"* property
- wikipedia pageviews to find popular (recognizable) images.

We also **replace linked text** in tables with images from Wikipedia URLs.

We **replace images** in up to **75% of a column** and **exclude tables with < 30% images** per column.



Team	P	W	T
Sweden	5	5	0
Denmark	5	3	1

Extract images corresponding to links
WIKIDATA
Filtering out categories to be replaced
Links to tables
Table HTML
Image Replacement

Team	P	W	T
	5	5	0
	5	3	1

MMTabQA Question Types

Explicit: Directly mention an entity replaced by an image.

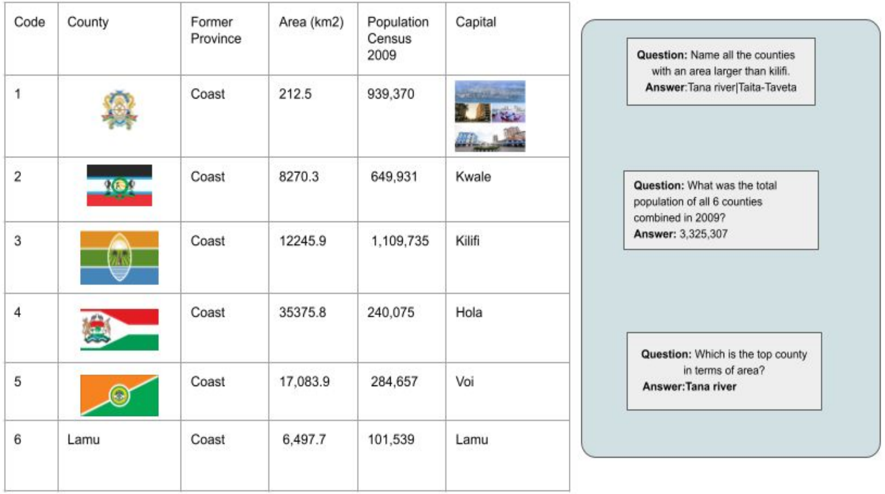
Answer-Mention: Answer includes an image-replaced entity, but the question does not.

Implicit: Use an image-replaced entity in reasoning without mentioning it in QA.

Visual: Focus on visual aspects, limited to specific categories like logos and flags.

- **Image Prompting:** VLM (Gemini 1.0) identifies category-specific attributes for images.

- **Attribute Replacement:** LLM (Gemini) provides and replaces attributes for explicit entity images in questions.



Code	County	Former Province	Area (km2)	Population Census 2009	Capital
1		Coast	212.5	939,370	
2		Coast	8270.3	649,931	Kwale
3		Coast	12245.9	1,109,735	Kilifi
4		Coast	35375.8	240,075	Hola
5		Coast	17,083.9	284,657	Voi
6	Lamu	Coast	6,497.7	101,539	Lamu

Question: Name all the counties with an area larger than kilifi.
Answer: Tana river|Taita-Taveta

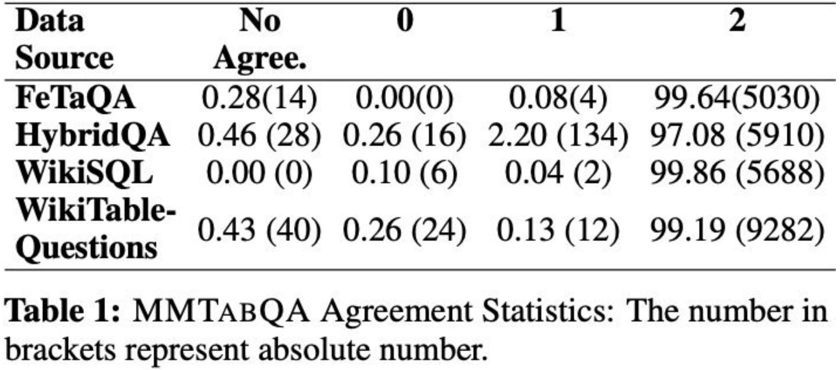
Question: What was the total population of all 6 counties combined in 2009?
Answer: 3,325,307

Question: Which is the top county in terms of area?
Answer:Tana river

MMTabQA Validation

Table Validation: We verify the correctness of entity replacements in recast tables.

- We sample 250 tables from each data source, totaling 750 tables.
- Three annotators rates all unique (image, original_text) pairs / table.



Data Source	No Agree.	0	1	2
FeTaQA	0.28(14)	0.00(0)	0.08(4)	99.64(5030)
HybridQA	0.46 (28)	0.26 (16)	2.20 (134)	97.08 (5910)
WikiSQL	0.00 (0)	0.10 (6)	0.04 (2)	99.86 (5688)
WikiTable-Questions	0.43 (40)	0.26 (24)	0.13 (12)	99.19 (9282)

Table 1: MMTabQA Agreement Statistics: The number in brackets represent absolute number.

Label 0 ⇒ incorrect image for the entity.

Label 1 ⇒ image represents the entity but is ambiguous.

Label 2 ⇒ image clearly represents the entity.

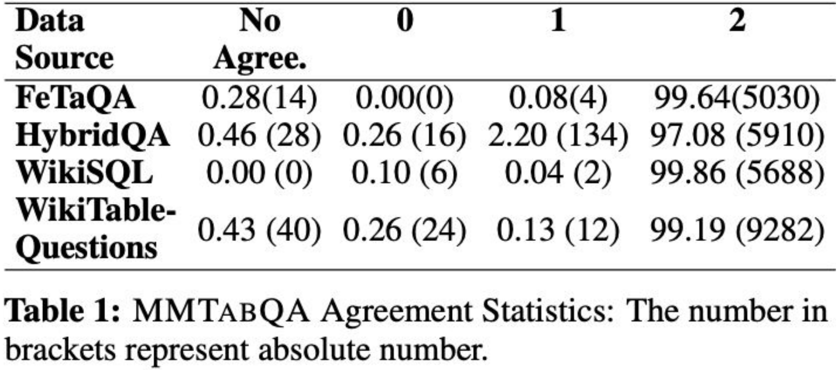
MMTabQA Validation

Table Validation: We verify the correctness of entity replacements in recast tables.

- We sample 250 tables from each data source, totaling 750 tables.
- Three annotators rates all unique (image, original_text) pairs / table.

Question Validation:

- *Repurposed Questions:* Explicit, answer-mention, and implicit questions from validated datasets.
- *Visual Synthetic Questions:* Require validation rating 500 recast questions
  - 21.46% scored as incorrect
  - 78.54% scored as appropriate.



Data Source	No Agree.	0	1	2
FeTaQA	0.28(14)	0.00(0)	0.08(4)	99.64(5030)
HybridQA	0.46 (28)	0.26 (16)	2.20 (134)	97.08 (5910)
WikiSQL	0.00 (0)	0.10 (6)	0.04 (2)	99.86 (5688)
WikiTable-Questions	0.43 (40)	0.26 (24)	0.13 (12)	99.19 (9282)

Table 1: MMTABQA Agreement Statistics: The number in brackets represent absolute number.

Label 0 ⇒ incorrect image for the entity.

Label 1 ⇒ image represents the entity but is ambiguous.

Label 2 ⇒ image clearly represents the entity.

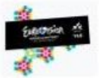
MMTabQA Examples

Code	County	Former Province	Area (km2)	Population Census 2009	Capital
1		Coast	212.5	939,370	
2		Coast	8270.3	649,931	Kwale
3		Coast	12245.9	1,109,735	Kilifi
4		Coast	35375.8	240,075	Hola
5		Coast	17,083.9	284,657	Voi
6	Lamu	Coast	6,497.7	101,539	Lamu

Question: Name all the counties with an area larger than kilifi.
Answer: Tana river\|Taita-Taveta

Question: What was the total population of all 6 counties combined in 2009?
Answer: 3,325,307

Question: Which is the top county in terms of area?
Answer:Tana river

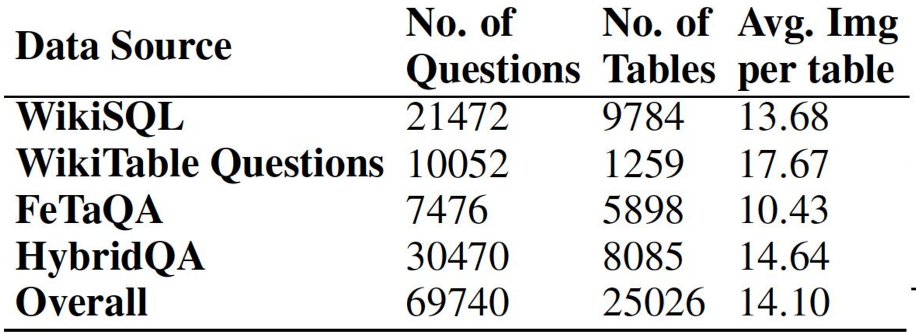
Year	Song	Artist	Place	Points	Composer
2002	Northern Girl	Prime Minister	10	55	
			2 (SF: 3rd)	248 (SF: 217)	
		Dmitry Koldun	6 (SF: 4th)	145 (SF: 176)	
	Peace Will come		11 (SF: 5)	83 (SF: 107)	

Question: Who is the Composer of \" peace will come \"?

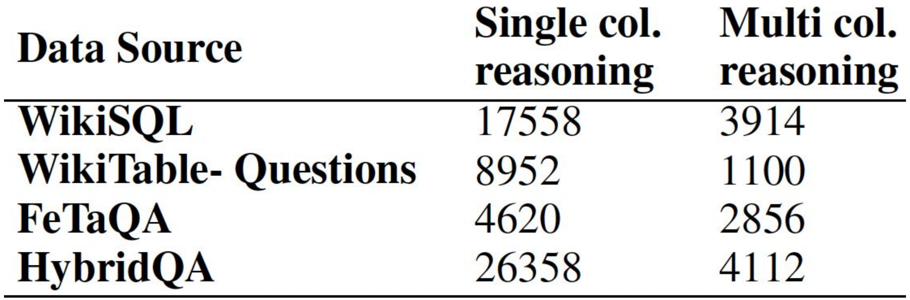
Answer: Kim Breitburg

Question: What is the place for the song \" work your magic \" in the event whose logo has a blue, green and pink color and has flower and star in it?
Answer: 6 (sf: 4th)

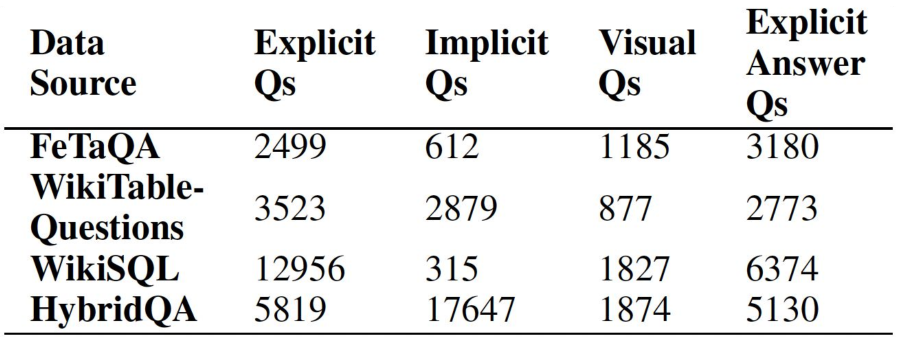
MMTabQA Statistics



Data Source	No. of Questions	No. of Tables	Avg. Img per table
WikiSQL	21472	9784	13.68
WikiTable Questions	10052	1259	17.67
FeTaQA	7476	5898	10.43
HybridQA	30470	8085	14.64
Overall	69740	25026	14.10



Data Source	Single col. reasoning	Multi col. reasoning
WikiSQL	17558	3914
WikiTable- Questions	8952	1100
FeTaQA	4620	2856
HybridQA	26358	4112

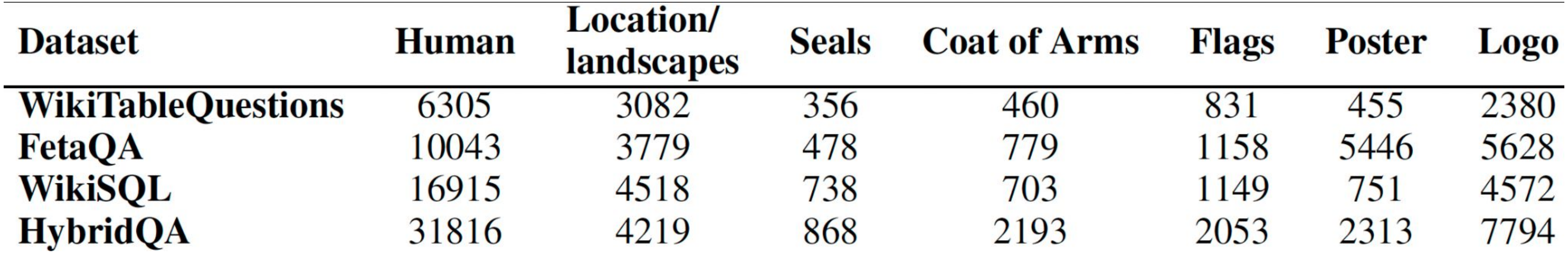


Data Source	Explicit Qs	Implicit Qs	Visual Qs	Explicit Answer Qs
FeTaQA	2499	612	1185	3180
WikiTable-Questions	3523	2879	877	2773
WikiSQL	12956	315	1827	6374
HybridQA	5819	17647	1874	5130

refer paper for other statistics

MMTabQA Statistics

Large **69,740** questions across **25,026** tables, featuring a diverse range of images.



Dataset	**Human**	**Location/ landscapes**	**Seals**	**Coat of Arms**	**Flags**	**Poster**	**Logo**
WikiTableQuestions	6305	3082	356	460	831	455	2380
FetaQA	10043	3779	478	779	1158	5446	5628
WikiSQL	16915	4518	738	703	1149	751	4572
HybridQA	31816	4219	868	2193	2053	2313	7794

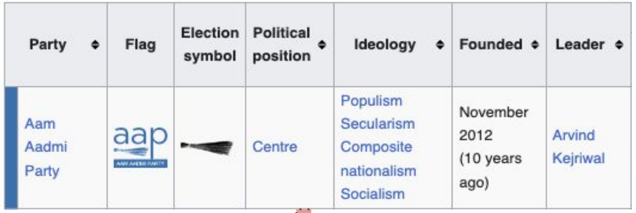
Partial Input Baseline

Images are excluded; image tags are replaced with **placeholders**.

The model **guesses entities** from image tags for question answering.

Acts as a **lower bound for performance** evaluation benchmark.

Models with direct image access should outperform this baseline.



Party	Flag	Election symbol	Political position	Ideology	Founded	Leader
Aam Aadmi Party			Centre	Populism Secularism Composite nationalism Socialism	November 2012 (10 years ago)	Arvind Kejriwal

Party	Flag	Election symbol	Political Position	Ideology	Founded	Leader
Aam Aadmi Party	<name entity>	<name entity>	Centre	Populism Secularism Composite nationalism Socialism	November 2012 (10 years ago)	Arvind Kejriwal

Party	Flag	Election symbol	Political Position	Ideology	Founded	Leader
Aam Aadmi Party	AAP	AAP	Centre	Populism Secularism Composite nationalism Socialism	November 2012 (10 years ago)	Arvind Kejriwal

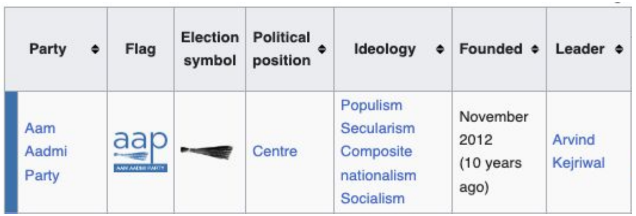
Dataset	WikiTableQuestions				WikiSQL				FetaQA			
Model	EQ	AQ	IQ	VQ	EQ	AQ	IQ	VQ	EQ	AQ	IQ	VQ
Partial Input Baseline												
Gemini-1.5 Flash	40.99	27.38	48.95	31.4	39.14	28.71	62.22	28	0.51	0.44	0.44	0.47
GPT-4o	57.45	38.02	70.83	42.40	52.57	43.86	72.38	39.00	0.51	0.46	0.42	0.44
Llama-3 70B	41.13	26.48	43.75	31.8	41.117	30.75	61.27	30.6	0.52	0.46	0.45	0.48
Mixtral 8x7B	26.56	9.90	30.26	20.2	23.42	17.71	28.88	19.2	0.44	0.39	0.38	0.39

Image-captioning Baseline

Converts tables to text-only using captions generated by VLMs.

Entity Prediction:

- VLMs predicts entities for the images infobox-style tables.
- VLMs generate caption and visual descriptions for each and every image.



Party	Flag	Election symbol	Political position	Ideology	Founded	Leader
Aam Aadmi Party			Centre	Populism Secularism Composite nationalism Socialism	November 2012 (10 years ago)	Arvind Kejriwal

Party	Flag	Election symbol	Political Position	Ideology	Founded	Leader
Aam Aadmi Party	aap broom blue Aam Aadmi Party	black broom lying horizontally	Centre	Populism Secularism Composite nationalism Socialism	November 2012 (10 years ago)	Arvind Kejriwal

Dataset	WikiTableQuestions				WikiSQL				FetaQA			
Model	EQ	AQ	IQ	VQ	EQ	AQ	IQ	VQ	EQ	AQ	IQ	VQ
Image-Captioning Baseline												
Gemini-1.5 Flash	52.34	42.16	51.39	42.2	50.42	40.85	67.30	46.6	0.57	0.46	0.42	0.43

Image-captioning Baseline

Converts tables to text-only using captions generated by VLMs.

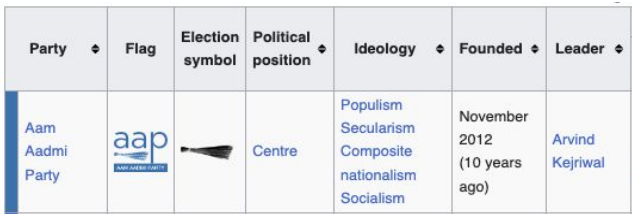
Entity Prediction:

- VLMs predicts entities for the images infobox-style tables.
- VLMs generate caption and visual descriptions for each and every image.

Question Answering:

- LLMs uses tables, questions, predicted entities, and visual descriptions to generate answers.
- Perform better with visual information, including captions and descriptions, especially for visual questions.

High computational cost due to image captioning, transforms task into text-only for LLM processing.

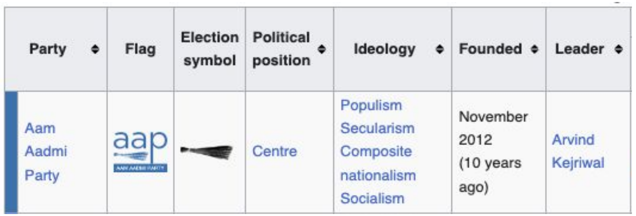


Party	Flag	Election symbol	Political position	Ideology	Founded	Leader
Aam Aadmi Party			Centre	Populism Secularism Composite nationalism Socialism	November 2012 (10 years ago)	Arvind Kejriwal

Party	Flag	Election symbol	Political Position	Ideology	Founded	Leader
Aam Aadmi Party	aap broom blue Aam Aadmi Party	black broom lying horizontally	Centre	Populism Secularism Composite nationalism Socialism	November 2012 (10 years ago)	Arvind Kejriwal

Dataset	WikiTableQuestions				WikiSQL				FetaQA			
Model	EQ	AQ	IQ	VQ	EQ	AQ	IQ	VQ	EQ	AQ	IQ	VQ
Image-Captioning Baseline												
Gemini-1.5 Flash	52.34	42.16	51.39	42.2	50.42	40.85	67.30	46.6	0.57	0.46	0.42	0.43

Table-Image Baseline

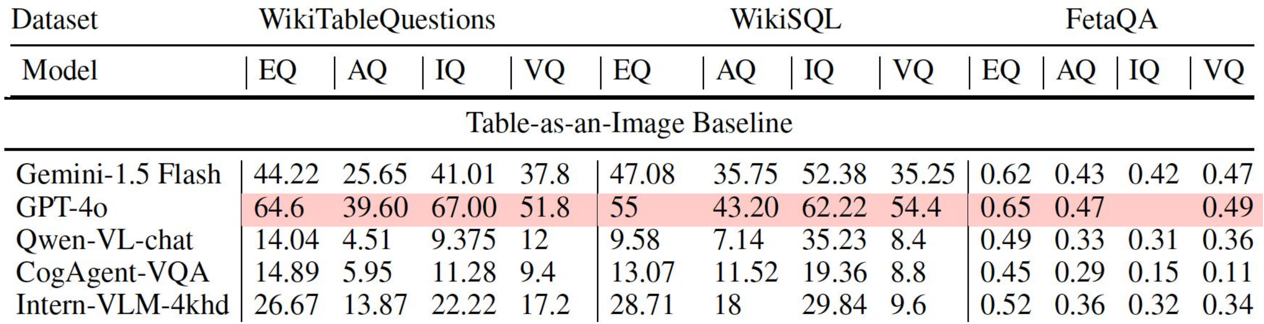


Party	Flag	Election symbol	Political position	Ideology	Founded	Leader
Aam Aadmi Party			Centre	Populism Secularism Composite nationalism Socialism	November 2012 (10 years ago)	Arvind Kejriwal

Creates an **image of the entire table,** including the entity images

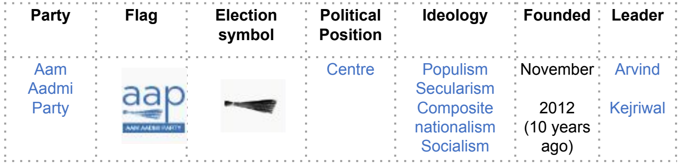
Inputs both the table image and textual question into the model.

Combines visual and textual modalities directly a.k.a **screenshot**



Dataset	WikiTableQuestions				WikiSQL				FetaQA			
Model	EQ	AQ	IQ	VQ	EQ	AQ	IQ	VQ	EQ	AQ	IQ	VQ
Table-as-an-Image Baseline												
Gemini-1.5 Flash	44.22	25.65	41.01	37.8	47.08	35.75	52.38	35.25	0.62	0.43	0.42	0.47
GPT-4o	64.6	39.60	67.00	51.8	55	43.20	62.22	54.4	0.65	0.47		0.49
Qwen-VL-chat	14.04	4.51	9.375	12	9.58	7.14	35.23	8.4	0.49	0.33	0.31	0.36
CogAgent-VQA	14.89	5.95	11.28	9.4	13.07	11.52	19.36	8.8	0.45	0.29	0.15	0.11
Intern-VLM-4khd	26.67	13.87	22.22	17.2	28.71	18	29.84	9.6	0.52	0.36	0.32	0.34

Interleaved Image-text Baseline



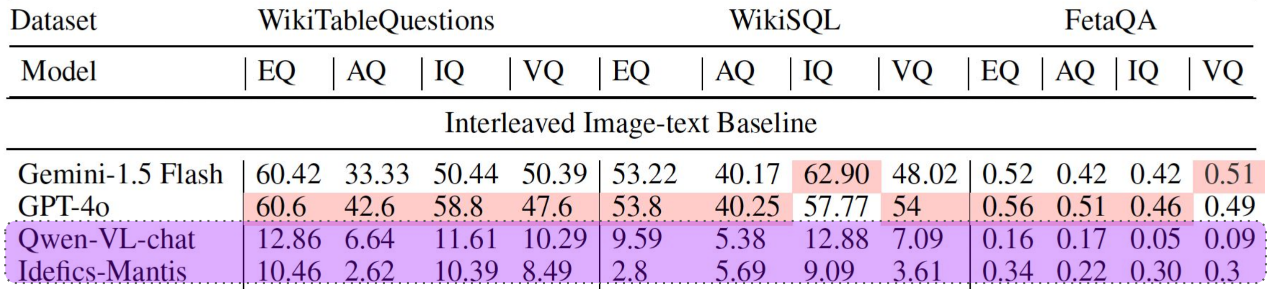
Party	Flag	Election symbol	Political Position	Ideology	Founded	Leader
Aam Aadmi Party			Centre	Populism Secularism Composite nationalism Socialism	November 2012 (10 years ago)	Arvind Kejriwal

Integrates both visual and textual modalities comprehensively together.

Unlike prior approaches combines modalities for optimal representation.

Balances visual & textual inputs for improved text-visual understanding.

Use LLMs for row pruning for efficiency on the open source models.



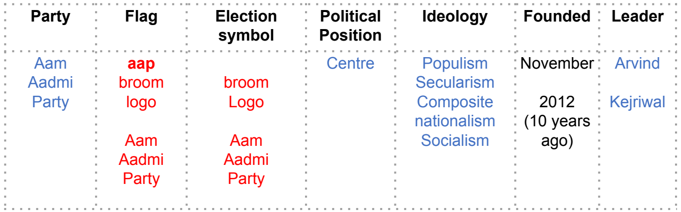
Dataset	WikiTableQuestions				WikiSQL				FetaQA			
Model	EQ	AQ	IQ	VQ	EQ	AQ	IQ	VQ	EQ	AQ	IQ	VQ
Interleaved Image-text Baseline												
Gemini-1.5 Flash	60.42	33.33	50.44	50.39	53.22	40.17	62.90	48.02	0.52	0.42	0.42	0.51
GPT-4o	60.6	42.6	58.8	47.6	53.8	40.25	57.77	54	0.56	0.51	0.46	0.49
Qwen-VL-chat	12.86	6.64	11.61	10.29	9.59	5.38	12.88	7.09	0.16	0.17	0.05	0.09
Idefics-Mantis	10.46	2.62	10.39	8.49	2.8	5.69	9.09	3.61	0.34	0.22	0.30	0.3

Oracle-Entity Replaced Baseline

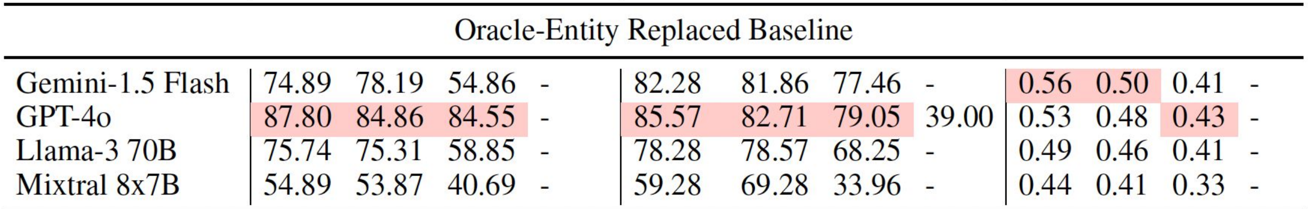
Evaluates textual tables where entities are replaced by original ones from the multimodal tables.

Sets an upper bound for performance in explicit, answer-mention, and implicit questions.

Not suitable for visual questions & assumes perfect entity disambiguation for tabular reasoning.



Party	Flag	Election symbol	Political Position	Ideology	Founded	Leader
Aam Aadmi Party	**aap** broom logo Aam Aadmi Party	broom Logo Aam Aadmi Party	Centre	Populism Secularism Composite nationalism Socialism	November 2012 (10 years ago)	Arvind Kejriwal



Oracle-Entity Replaced Baseline

Gemini-1.5 Flash	74.89	78.19	54.86	-	82.28	81.86	77.46	-	0.56	0.50	0.41	-
GPT-4o	87.80	84.86	84.55	-	85.57	82.71	79.05	39.00	0.53	0.48	0.43	-
Llama-3 70B	75.74	75.31	58.85	-	78.28	78.57	68.25	-	0.49	0.46	0.41	-
Mixtral 8x7B	54.89	53.87	40.69	-	59.28	69.28	33.96	-	0.44	0.41	0.33	-

Full Results

Results on sampled subset.

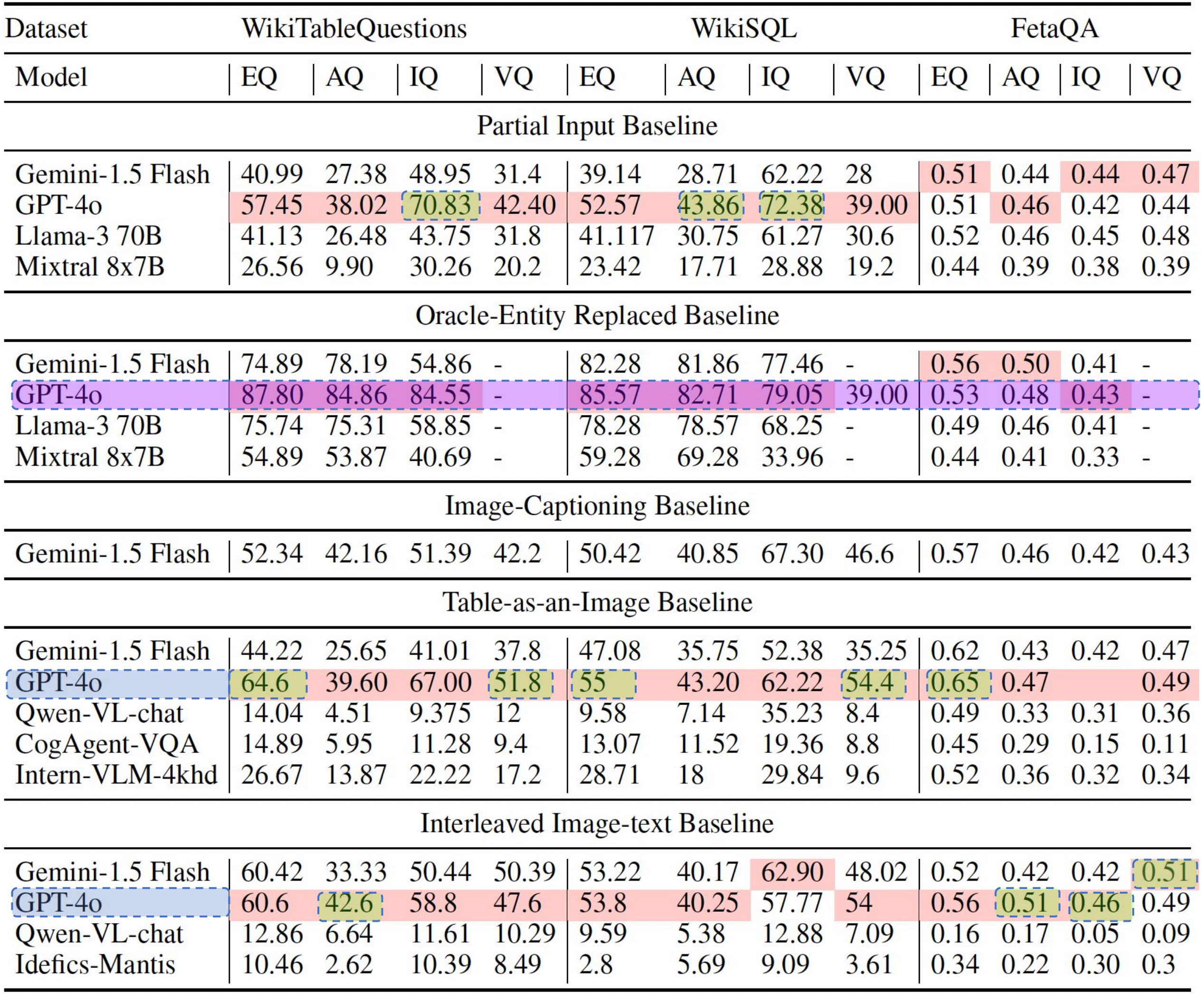
Substring match for Wiki-related data sources

ROUGE-L for FetaQA data source.

Best model performances are highlighted in Red

EQ - Explicit,
AQ - Answer-Mention
IQ - Implicit
VQ -Visual

Best Performance



Dataset	WikiTableQuestions				WikiSQL				FetaQA			
Model	EQ	AQ	IQ	VQ	EQ	AQ	IQ	VQ	EQ	AQ	IQ	VQ
Partial Input Baseline												
Gemini-1.5 Flash	40.99	27.38	48.95	31.4	39.14	28.71	62.22	28	0.51	0.44	0.44	0.47
GPT-4o	57.45	38.02	70.83	42.40	52.57	43.86	72.38	39.00	0.51	0.46	0.42	0.44
Llama-3 70B	41.13	26.48	43.75	31.8	41.117	30.75	61.27	30.6	0.52	0.46	0.45	0.48
Mixtral 8x7B	26.56	9.90	30.26	20.2	23.42	17.71	28.88	19.2	0.44	0.39	0.38	0.39
Oracle-Entity Replaced Baseline												
Gemini-1.5 Flash	74.89	78.19	54.86	-	82.28	81.86	77.46	-	0.56	0.50	0.41	-
GPT-4o	87.80	84.86	84.55	-	85.57	82.71	79.05	39.00	0.53	0.48	0.43	-
Llama-3 70B	75.74	75.31	58.85	-	78.28	78.57	68.25	-	0.49	0.46	0.41	-
Mixtral 8x7B	54.89	53.87	40.69	-	59.28	69.28	33.96	-	0.44	0.41	0.33	-
Image-Captioning Baseline												
Gemini-1.5 Flash	52.34	42.16	51.39	42.2	50.42	40.85	67.30	46.6	0.57	0.46	0.42	0.43
Table-as-an-Image Baseline												
Gemini-1.5 Flash	44.22	25.65	41.01	37.8	47.08	35.75	52.38	35.25	0.62	0.43	0.42	0.47
GPT-4o	64.6	39.60	67.00	51.8	55	43.20	62.22	54.4	0.65	0.47		0.49
Qwen-VL-chat	14.04	4.51	9.375	12	9.58	7.14	35.23	8.4	0.49	0.33	0.31	0.36
CogAgent-VQA	14.89	5.95	11.28	9.4	13.07	11.52	19.36	8.8	0.45	0.29	0.15	0.11
Intern-VLM-4khd	26.67	13.87	22.22	17.2	28.71	18	29.84	9.6	0.52	0.36	0.32	0.34
Interleaved Image-text Baseline												
Gemini-1.5 Flash	60.42	33.33	50.44	50.39	53.22	40.17	62.90	48.02	0.52	0.42	0.42	0.51
GPT-4o	60.6	42.6	58.8	47.6	53.8	40.25	57.77	54	0.56	0.51	0.46	0.49
Qwen-VL-chat	12.86	6.64	11.61	10.29	9.59	5.38	12.88	7.09	0.16	0.17	0.05	0.09
Idefics-Mantis	10.46	2.62	10.39	8.49	2.8	5.69	9.09	3.61	0.34	0.22	0.30	0.3

Across Approaches Comparisons

Partial Input Baseline

- **Lower Bound:** Sets the lower limit for performance.
- **Expectation:** Models with image access should perform better.

Oracle-Entity Replaced Baseline

- **Upper Bound:** Sets the upper limit for performance.
- **Ideal Conditions:** Assumes perfect entity disambiguation.

Table-as-image Baseline

- **Improvement:** Better than Partial Input baseline.
- **Limitation:** Struggles with interpreting table structure from images.

Interleaved Image-text Baseline

- **Enhanced Performance:** Better than Table-as-image baseline.
- **Limitation:** Does not reach the upper bound performance.

Image-as-caption Approach

- **Moderate Performance:** Better than Table-as-image baseline.
- **Less Effective:** Not as good as Interleaved Image-text baseline.

Across Question Types

Explicit Questions:

Contain clear, specific entity mentions.

Achieve the highest parameter scores.

Implicit Questions:

Require additional reasoning to infer answers.

lower performance compared to explicit questions.

struggle as require multi-step or intricate reasoning

Explicit Answer Mention Questions:

Challenge due to the need for precise entity disambiguation.

marked decrease in performance, in "Table as an Image" approach.

Visual Questions:

Perform better than answer-mention questions.

Require visual understanding of entities in images.

Show lower performance metrics (37.8% on WikiSQL, 35.25% on WTQ).

refer paper for other analysis

Across Data-QA Types

Explicit Questions:

Contain clear, specific entity mentions.

Achieve the highest parameter scores.

Implicit Questions:

Require additional reasoning to infer answers.

lower performance compared to explicit questions.

struggle as require multi-step or intricate reasoning

Explicit Answer Mention Questions:

Challenge due to the need for precise entity disambiguation.

marked decrease in performance, in "Table as an Image" approach.

Visual Questions:

Perform better than answer-mention questions.

Require visual understanding of entities in images.

Show lower performance metrics (37.8% on WikiSQL, 35.25% on WTQ).

Contrast: WikiSQL vs. WikiTableQuestions:

Implicit questions perform better in WikiSQL (52.3% vs. 47.08%).

Explicit questions perform better in WikiTableQuestions (WTQ).

Anomalous FeTaQA Dataset:

Current metrics limit the extent of inferences.

Where do Model Fails?

- models often **misidentify entities** from the present images
    - struggle to recognize crucial visual elements
- difficulties **handling excessive (multiple images)** content
    - leading to incomplete or incorrect retrievals
- even when models correctly identify entities
    - can still reach **incorrect conclusions (reasoning)**
- open-source underperform compared to closed-source ones
    - drop performance as image complexity increases.
    - resulting in inaccurate interpretation & limited reasoning abilities
    - possibly models often lack robust vision encoders

enhancing VLMs with MMTabQA for real-world multimodal tables

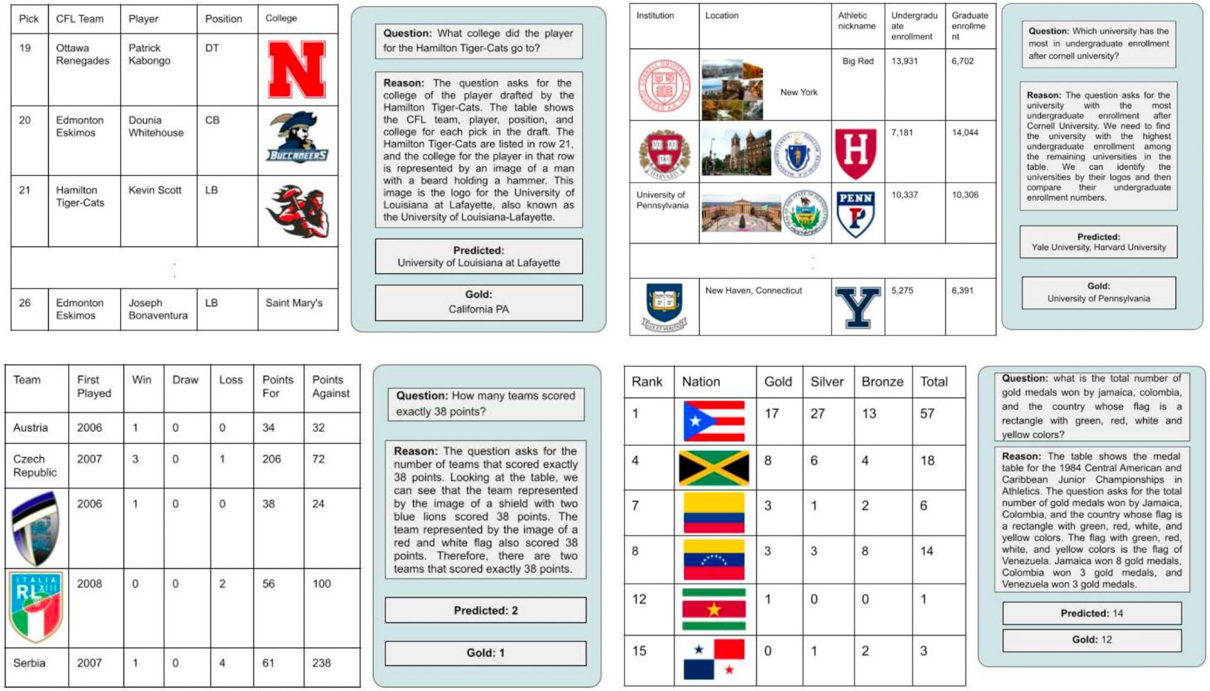
Takeaway

VLM Struggle

Entity Disambiguation within Table Contexts (Fig. 3(a)).

Concurrent Inference from Multiple Images (Fig. 3(b)).

Identification of Visual Attributes in Images (Fig. 3(c)).



Pick	CFL Team	Player	Position	College
19	Ottawa Renegades	Patrick Kabongo	DT	
20	Edmonton Eskimos	Dounia Whitehouse	CB	
21	Hamilton Tiger-Cats	Kevin Scott	LB	
⋮				
26	Edmonton Eskimos	Joseph Bonaventura	LB	Saint Mary's

Question: What college did the player for the Hamilton Tiger-Cats go to?

Reason: The question asks for the college of the player drafted by the Hamilton Tiger-Cats. The table shows the CFL team, player, position, and college for each pick in the draft. The Hamilton Tiger-Cats are listed in row 21, and the college for the player in that row is represented by an image of a man with a beard holding a hammer. This image is the logo for the University of Louisiana at Lafayette, also known as the University of Louisiana-Lafayette.

Predicted:
University of Louisiana at Lafayette

Gold:
California PA

Institution	Location	Athletic nickname	Undergraduate enrollment	Graduate enrollment
	New York	Big Red	13,931	6,702
			7,181	14,044
University of Pennsylvania			10,337	10,306
⋮				
	New Haven, Connecticut		5,275	6,391

Question: Which university has the most in undergraduate enrollment after cornell university?

Reason: The question asks for the university with the most undergraduate enrollment after Cornell University. We need to find the university with the highest undergraduate enrollment among the remaining universities in the table. We can identify the universities by their logos and then compare their undergraduate enrollment numbers.

Predicted:
Yale University, Harvard University

Gold:
University of Pennsylvania

Team	First Played	Win	Draw	Loss	Points For	Points Against
Austria	2006	1	0	0	34	32
Czech Republic	2007	3	0	1	206	72
	2006	1	0	0	38	24
	2008	0	0	2	56	100
Serbia	2007	1	0	4	61	238

Question: How many teams scored exactly 38 points?

Reason: The question asks for the number of teams that scored exactly 38 points. Looking at the table, we can see that the team represented by the image of a shield with two blue lions scored 38 points. The team represented by the image of a red and white flag also scored 38 points. Therefore, there are two teams that scored exactly 38 points.

Predicted: 2

Gold: 1

Rank	Nation	Gold	Silver	Bronze	Total
1		17	27	13	57
4		8	6	4	18
7		3	1	2	6
8		3	3	8	14
12		1	0	0	1
15		0	1	2	3

Question: what is the total number of gold medals won by jamaica, colombia, and the country whose flag is a rectangle with green, red, white and yellow colors?

Reason: The table shows the medal table for the 1984 Central American and Caribbean Junior Championships in Athletics. The question asks for the total number of gold medals won by Jamaica, Colombia, and the country whose flag is a rectangle with green, red, white, and yellow colors. The flag with green, red, white, and yellow colors is the flag of Venezuela. Jamaica won 8 gold medals, Colombia won 3 gold medals, and Venezuela won 3 gold medals.

Predicted: 14

Gold: 12

Figure 3: Clockwise from left - (a): Table about College Football, (b): Table about College Enrollment, (c): Table about 1984 Central American Games, (d): Table about International Football.

APPENDIX

Pick	CFL Team	Player	Position	College
19	Ottawa Renegades	Patrick Kabongo	DT	
20	Edmonton Eskimos	Dounia Whitehouse	CB	
21	Hamilton Tiger-Cats	Kevin Scott	LB	
		:		
26	Edmonton Eskimos	Joseph Bonaventura	LB	Saint Mary's

Question: What college did the player for the Hamilton Tiger-Cats go to?

Reason: The question asks for the college of the player drafted by the Hamilton Tiger-Cats. The table shows the CFL team, player, position, and college for each pick in the draft. The Hamilton Tiger-Cats are listed in row 21, and the college for the player in that row is represented by an image of a man with a beard holding a hammer. This image is the logo for the University of Louisiana at Lafayette, also known as the University of Louisiana-Lafayette.

Predicted:
University of Louisiana at Lafayette

Gold:
California PA

Institution	Location	Athletic nickname	Undergraduate enrollment	Graduate enrollment
	New York	Big Red	13,931	6,702
			7,181	14,044
University of Pennsylvania			10,337	10,306
	:			
	New Haven, Connecticut		5,275	6,391

Question: Which university has the most in undergraduate enrollment after cornell university?

Reason: The question asks for the university with the most undergraduate enrollment after Cornell University. We need to find the university with the highest undergraduate enrollment among the remaining universities in the table. We can identify the universities by their logos and then compare their undergraduate enrollment numbers.

Predicted:
Yale University, Harvard University

Gold:
University of Pennsylvania

Team	First Played	Win	Draw	Loss	Points For	Points Against
Austria	2006	1	0	0	34	32
Czech Republic	2007	3	0	1	206	72
	2006	1	0	0	38	24
	2008	0	0	2	56	100
Serbia	2007	1	0	4	61	238

Question: How many teams scored exactly 38 points?

Reason: The question asks for the number of teams that scored exactly 38 points. Looking at the table, we can see that the team represented by the image of a shield with two blue lions scored 38 points. The team represented by the image of a red and white flag also scored 38 points. Therefore, there are two teams that scored exactly 38 points.

Predicted: 2

Gold: 1

Rank	Nation	Gold	Silver	Bronze	Total
1		17	27	13	57
4		8	6	4	18
7		3	1	2	6
8		3	3	8	14
12		1	0	0	1
15		0	1	2	3

Question: what is the total number of gold medals won by jamaica, colombia, and the country whose flag is a rectangle with green, red, white and yellow colors?

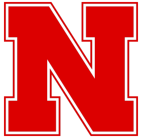
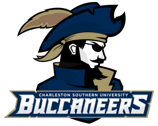
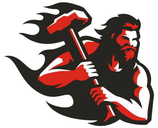
Reason: The table shows the medal table for the 1984 Central American and Caribbean Junior Championships in Athletics. The question asks for the total number of gold medals won by Jamaica, Colombia, and the country whose flag is a rectangle with green, red, white, and yellow colors. The flag with green, red, white, and yellow colors is the flag of Venezuela. Jamaica won 8 gold medals, Colombia won 3 gold medals, and Venezuela won 3 gold medals.

Predicted: 14

Gold: 12

Error Analysis

Incorrect Entity Disambiguation

Pick	CFL Team	Player	Position	College
19	Ottawa Renegades	Patrick Kabongo	DT	
20	Edmonton Eskimos	Dounia Whitehouse	CB	
21	Hamilton Tiger-Cats	Kevin Scott	LB	
		. .		
26	Edmonton Eskimos	Joseph Bonaventura	LB	Saint Mary's

Question: What college did the player for the Hamilton Tiger-Cats go to?

Reason: The question asks for the college of the player drafted by the Hamilton Tiger-Cats. The table shows the CFL team, player, position, and college for each pick in the draft. The Hamilton Tiger-Cats are listed in row 21, and the college for the player in that row is represented by an image of a man with a beard holding a hammer. This image is the logo for the University of Louisiana at Lafayette, also known as the University of Louisiana-Lafayette.

Predicted:
University of Louisiana at Lafayette

Gold:
California PA

Error Analysis

Incorrect Reasoning

Institution	Location	Athletic nickname	Undergraduate enrollment	Graduate enrollment
	New York	Big Red	13,931	6,702
			7,181	14,044
University of Pennsylvania			10,337	10,306
. .				
	New Haven, Connecticut		5,275	6,391

Question: Which university has the most in undergraduate enrollment after cornell university?

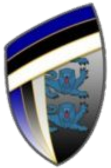
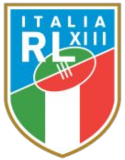
Reason: The question asks for the university with the most undergraduate enrollment after Cornell University. We need to find the university with the highest undergraduate enrollment among the remaining universities in the table. We can identify the universities by their logos and then compare their undergraduate enrollment numbers.

Predicted:
Yale University, Harvard University

Gold:
University of Pennsylvania

Error Analysis

Incorrect Retrieval

Team	First Played	Win	Draw	Loss	Points For	Points Against
Austria	2006	1	0	0	34	32
Czech Republic	2007	3	0	1	206	72
	2006	1	0	0	38	24
	2008	0	0	2	56	100
Serbia	2007	1	0	4	61	238

Question: How many teams scored exactly 38 points?

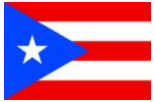
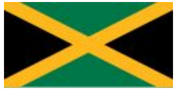
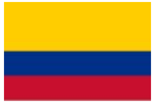
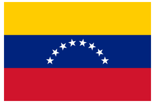
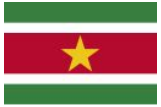
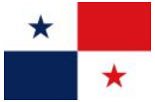
Reason: The question asks for the number of teams that scored exactly 38 points. Looking at the table, we can see that the team represented by the image of a shield with two blue lions scored 38 points. The team represented by the image of a red and white flag also scored 38 points. Therefore, there are two teams that scored exactly 38 points.

Predicted: 2

Gold: 1

Error Analysis

Incorrect Visual Reasoning

Rank	Nation	Gold	Silver	Bronze	Total
1		17	27	13	57
4		8	6	4	18
7		3	1	2	6
8		3	3	8	14
12		1	0	0	1
15		0	1	2	3

Question: what is the total number of gold medals won by jamaica, colombia, and the country whose flag is a rectangle with green, red, white and yellow colors?

Reason: The table shows the medal table for the 1984 Central American and Caribbean Junior Championships in Athletics. The question asks for the total number of gold medals won by Jamaica, Colombia, and the country whose flag is a rectangle with green, red, white, and yellow colors. The flag with green, red, white, and yellow colors is the flag of Venezuela. Jamaica won 8 gold medals, Colombia won 3 gold medals, and Venezuela won 3 gold medals.

Predicted: 14

Gold: 12

Multimodal Image-Text Processing

- Previously, NLP and Computer Vision were considered challenging fields in Computer Science.
- The advent of Transformer Architecture and models like BERT, ViT, and GPT revolutionized both fields.

- This advancement paved the way for combining both fields into **Multimodal Image-Text Processing**.
- Key tasks in this domain include Image Captioning, **Visual Question Answering**, Image Retrieval, Image-Text Sentiment Analysis, and Multimodal Summarization.

Large Language Models

- LSTMs and ELMo were state-of-the-art until 2017 when the Transformer architecture led to the development of first-generation LLMs like BERT and GPT.

- These models showcased emergent abilities such as zero-shot learning, in-context learning, and Chain of Thought (CoT) reasoning.

- Current state-of-the-art LLMs include GPT-4, LLaMa 3, Gemini.

Large Language Models

- Multimodal Transformers like CLIP, Visual BERT, and ViLT process text, images, audio, and point cloud data, inspired by NLP models.

OpenAI
CLIP
CONNECTING TEXT AND IMAGES

- Recent MLLMs like GPT-4V and Gemini-Vision expand LLM capabilities across multiple modalities, building on models such as GPT-3.5.

- Our work focuses on experiments with the **Gemini-1.0-Pro** family of models due to accessibility and resource constraints with other LLMs/MLLMs.

LAVIS